RICE UNIVERSITY

Local Sociophonetic Knowledge in Speech Perception

by

Christian Koops

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Nancy Niedzielski, Associate Professor,
Chair, Linguistics

Katherine Crosswhite, Assistant Professor
Linguistics

Stephen L. Klineberg, Professor
Sociology

HOUSTON, TEXAS
APRIL 2011

ABSTRACT

Local Sociophonetic Knowledge in Speech Perception

by

Christian Koops

Sociophonetic studies of speech perception have demonstrated that the social identity which listeners attribute to a speaker can lead to predictable biases in the way speech sounds produced by that speaker are linguistically categorized (e.g., Strand & Johnson 1996; Niedzielski 1999; Hay, Warren & Drager 2006). This has been observed where listeners use available social information about a speaker to resolve lexical ambiguity. However, less is known about the role of sociophonetic knowledge in speech perception when listeners are not faced with global linguistic ambiguity. Drawing on Strand's (2000) study of the processing effects of gender typicality, this dissertation investigates whether sociophonetic knowledge can facilitate or inhibit unambiguous spoken word recognition. Based

on a survey of sociophonetic variation in the Houston metropolitan area, predictions are formulated for the processing of words containing four vowels: /eɪ/ and /ɛ/ in the speech of older and younger Anglos, and /ɑ/ and /ʌ/ in the speech of young Anglos and young African-Americans. Houston listeners identified words containing variants of these vowels in a congruent condition and in an incongruent condition. In the congruent condition the combination of speaker identity and vowel variant was designed to match the listener's knowledge of local language variation. In the incongruent condition, it was designed to contradict it. A congruency effect was found for some but not all vowels. The results indicate that social information about a speaker can also affect speech perception in the absence of lexical ambiguity, but only where words are at least temporarily ambiguous. Where there is no linguistic ambiguity at all, perception can be unaffected by sociophonetic knowledge. These results are discussed in the context of Luce, McLennan & Charles-Luce's (2003) time course hypothesis and in the context of exemplar-based models of sociophonetic knowledge (Johnson 1997, Pierrehumbert 2001).

ACKNOWLEDGMENTS

Many thanks are in order. To begin, I owe a tremendous debt of gratitude to my family, especially to my parents Helga and Erwin Koops for their unconditional and unwavering support of my choice to become as a linguist.

Next, I thank the three members of my thesis committee for their help and criticism in this research and previous work that gave rise to it. First and foremost, I am still occasionally in disbelief at how lucky I have been to have Nancy Niedzielski as my thesis advisor. Nancy, you not only introduced me to sociolinguistics, phonetics, and sociophonetics, but you are also the most inspiring, dependable, and good-humored advisor anyone in grad school can hope for. I have learned a tremendous amount from Katherine Crosswhite, who introduced me to laboratory phonology, speech and hearing science, Praat scripting and more, and who always, without exception, knew what the problem was when a script crashed or the stats didn't come out as expected. I also thank Stephen Klineberg for his advice and criticism and for, in the run-up to this

dissertation, allowing me to audit his SOCI 308 "Houston – The Sociology of a City" – without a doubt one of the most interesting classes I've been in at Rice.

More thanks go to the other current and former Rice Linguistics faculty members who have influenced and helped me, whether in sociolinguistics, semantics, typology, or otherwise: Michel Achard, Michael Barlow, Suzanne Kemmer, Matt Shibatani, Jim Stanford, and Christina Willis.

Finally, special thanks go to the many friends at Rice and beyond who have helped and inspired me or sometimes just been there, in one way or another. In chronological order, over the years: Tanja Kupisch, Amir Bar, Monica Sanaphre, Dave Katten, Caleb Everett, Chris Taylor, Martin Hilpert, Gu-Jing Lin, Jessica Hartstein, Sarah Lee, Vica Papp, Chris Schmidt, Nori Nagaya, Jayeon Jeong, Cassandra Pace, Elizabeth Brunner, Andrew Pantos, Mamadou Cissé, Ann Olivo, Jen Hoecker, Carlos Molina-Vital, Haowen Jiang, Penelope Howe, and Rachael Allbritten.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

1. Introduction

The topic of this dissertation is language users' knowledge of sociophonetic variation. I am using the term *sociophonetic variation* in the sense that it has in the field of variationist sociolinguistics (Labov 1966, 1994, 2001). Broadly speaking, it refers to alternative realizations, or variants, of a speech sound that are correlated with aspects of the speakers' social identity. For example, in a speech community where younger speakers tend to produce the most advanced variants of a speech sound undergoing sound change, say, the most fronted variants of the English vowel /u/, this phonetic variation and its social distribution together constitute a case of sociophonetic variation. As I discuss in detail in this chapter, what is at issue in the current study is not primarily the knowledge that speakers have of the phonetic variants they themselves use and

that allows them to produce them natively. Rather, I am concerned which the knowledge that language users have as listeners of the variation that exists in the speech of other speakers in their speech community. Throughout this dissertation, I will refer to such knowledge as *sociophonetic knowledge.*

I am interested in sociophonetic knowledge because such knowledge has been shown to inform the phonetic perception of speech (e.g., Strand & Johnson 1996). The research presented here builds on a growing body of experimental work carried out by sociolinguists and psycholinguistics which suggests that speech perception is finely tuned to the varying production of speech sounds that listeners experience in their local speech community. Listeners appear to make systematic use of their accumulated knowledge of sociophonetic variation when they linguistically interpret the speech of others (Niedzielski 1997, 1999; Drager 2005, 2011; Hay, Warren & Drager 2006; Staum 2008 *inter alia*). This is most clearly seen in the fact that listeners may interpret the same acoustic signal differently if that signal is understood to be spoken by different speakers. For instance, a speech sound may be perceived as instantiating different phonemic categories depending on whether the speaker is perceived to be a male speaker

or a female speaker (Strand & Johnson 1996; Johnson, Strand & D'Imperio 1999).

As I discuss in this chapter, the goal of most previous research on sociophonetic knowledge has been to determine how encompassing and differentiated such knowledge is. What types of social variation do listeners display awareness of, and how much of that is available to speech perception? As more results have been reported, it has become clear that the social knowledge which listeners access in categorizing speech sounds is remarkably detailed and often closely dovetails the findings of sociolinguistics working in the same speech communities (Drager 2005, Hay, Warren & Drager 2006; Koops, Gentry & Pantos 2008). It appears that sociophonetic knowledge includes knowledge of variation along the same social dimensions that are traditionally studied in sociolinguistics: variation in age, gender, regional origin, ethnicity, and social class.

However, less effort has been expended on exploring the question under exactly which conditions listeners adjust their perception to anticipate the speech of different speakers in this way and under which conditions they do not.

Prior experiments have demonstrated the role of sociophonetic knowledge in speech perception primarily by creating a particular type of processing condition. Typically the listeners' task in these studies was to disambiguate a globally ambiguous lexical item in one of two ways. Their decision as to which interpretation is the more likely one depended on whether a particular dialect feature was assumed to be present or absent in the perceived speaker's language variety. In this dissertation, I argue that as a result of choosing this methodology, it is not clear whether perceptual effects of sociophonetic knowledge are restricted to these particular processing conditions or whether sociophonetic knowledge affects speech perception more generally. Therefore, the goal of this dissertation is to determine how pervasive the role of sociophonetic knowledge is in speech perception. Specifically, does it also play a role in processing linguistic structures that are not globally ambiguous?

The remainder of this chapter is organized as follows. In Section 1.1, I begin the discussion of speech perception by reviewing prior findings that point to a strong correlation between the categories of speech production and the categories of speech perception. In Section 1.2, I discuss the concept of

sociophonetic knowledge in greater detail. I review a second set of studies dealing specifically with listeners' capacity to infer social information from speech cues. In Section 1.3, I review a series of studies which have aimed to determine whether, as sketched above, social cues are relevant to speech perception. In Section 1.4, I summarize the limitations of this series of studies and formulate a specific hypothesis with regard to the processing of sociophonetic variants to be tested in later chapters.

1.1. Production-perception correlations

Variationist sociolinguists have studied the perceptual side of sociophonetic variation primarily with regard to the question of cross-dialect perception and misperception (e.g., Labov, Yaeger & Steiner 1972, Labov & Ash 1997). The hypothesis underlying this research tradition is that differing speech norms may lead to misperception when speakers of one dialect hear phonetic variants that are characteristic of another dialect but not of their own. One cause of different speech norms that has received particular attention are ongoing regional vowel

shifts in North America such as those documented in detail by Labov, Yaeger & Steiner (1972). For example, Labov et al. (1972:135ff) report a speech perception experiment in which listeners in Philadelphia were asked to identify words containing fronted variants of /u/ produced by Coastal North Carolina speakers. The fronting of /u/ in these speakers results in a vowel that ends in an [i]-like quality. As predicted, the listeners from Philadelphia showed a tendency to misperceive the stimuli as words containing /i/.

A larger project devoted to the same question was reported in Labov & Ash (1997). The basis of this study were speech samples from three Anglo dialects: Chicago, Philadelphia, and Birmingham. Again, the samples used in the experiment were variants which represent the outcomes of regional sound changes, for example fronted /ɑ/ in the speech of the Chicago speakers and monophthongal /aɪ/ in the speech of the Birmingham speakers. As predicted, high rates of misperception were found for listeners from non-matching dialect areas. In fact, even listeners from the speakers' own dialect region showed imperfect recognition, especially when the word was presented in isolation. Still,

recognition rates by local listeners were generally higher than those of non-local listeners.

A similar effect of dialect-specific perception was demonstrated by Willis (1972) using synthetically produced vowels. Willis studied vowel perception by high school students from two adjacent cities across the US-Canada border, Fort Erie, Ontario, and Buffalo, New York. He showed, for example, that the position of the vowels /æ/ and /ɛ/ in the speech of members of each community is correlated with their perception of synthetic tokens of these vowels. The listeners from each community categorized vowel tokens as belonging to different phonemic categories in accordance with their own production norms. For example, in accordance with the more raised /æ/ in Buffalo, listeners from that city were more likely to categorize vowel qualities intermediate between [ɛ] and [æ] as /æ/ than the listeners from Fort Erie. Similar results were obtained by Janson (1983, 1986) for Swedish. Flanigan & Norris (2000) and Clopper, Pierrehumbert & Tamati (2008) report results of more recent cross-dialect misperception experiments in North America along similar theoretical lines.

Overall, the results of these experiments point to a clear correlation between the categories of speech production and speech perception in members of a speech community. It appears to be possible to predict, at least to some extent, how a given vowel quality will be categorized by listeners based on the way that the listeners themselves produce the relevant vowel. Another way of stating this is to say that listeners operate under the latent assumption that other speakers generally sound like they themselves do.

However, other findings complicate such a direct link between production on perception. One type of complicating evidence has come from the study of vowel merger, especially the study of what has come to be known as "near-merger" (Labov, Karen & Miller 1991). For example, Janson & Schulman (1983) investigated the perceptual consequences of the merger of short /e/ and /ɛ/ in two dialects of Swedish. Listeners from Stockholm, who no longer distinguish the two vowels, performed the same task as listeners from Northern Sweden, where the distinction is still maintained. Surprisingly, both listener groups were largely unable to distinguish short /e/ and /ɛ/. Labov, Karen & Miller (1991) replicated this effect in Philadelphia listeners where the variable in question was

the merger of /ɛ/ and /ʌ/ before /ɹ/, as in *merry* and *Murry*. The authors found that even Philadelphia speakers who maintained a distinction between the two vowels in production showed a severely limited ability to distinguish between them in perception (see also Bowie 2001a, 2001b).

A possible explanation for this apparent asymmetry between production and perception is that listeners' perceptual categories are not directly a function of their own production but rather the product of hearing the speech of others in their community. The experience of listening to, for example, Philadelphia speakers who do not maintain the relevant distinction and produce pre-rhotic /ɛ/ in a way that corresponds to the listener's own /ʌ/, and vice versa, leads these listeners to routinely ignore the contrast when listening to others before abandoning it in their own speech. The phenomenon of near-merger thereby suggests that listeners' perceptual categories are structured so as to best match the speech of interlocutors. The finding that there is, nevertheless, a general production-perception correlation can be ascribed to the fact that a listener's most frequent interlocutors are likely to be in-group speakers who produce variants that are very similar to the listener's own variants.

Further evidence for this more differentiated view of the production-perception correlation comes from Warren, Hay and Thomas' (2007) study of the perception of the merger of /iə/ and /eə/, as in the words *here* and *hair*, in New Zealand English. Here, the two vowels are increasingly being merged, and many young speakers no longer maintain the contrast. Warren et al. report that, surprisingly, some younger New Zealanders were able to distinguish /iə/ and /eə/ reliably in the speech of an older New Zealand English speaker despite reporting that they themselves do not maintain the distinction in their own speech. The authors explain this finding with the variable exposure that their participants have to merged and non-merged variants. They relate it to a correlation between the likelihood of merger and the speaker's social class. Previous production survey research had shown that speakers from higher social classes are less likely to merge the two vowels than speakers from lower social classes. Mirroring this finding, Warren et al. found that the listeners' discrimination rate also increased with their social class. Therefore, Warren et al. argue, it seems that younger speakers who hear more distinct variants in

their social environment will also be able to distinguish them more reliably, regardless of how they themselves produce the vowels.

In general, then, Warren et al.'s finding speaks to the idea that perceptual categories are primarily a product of long-term perceptual experience and not directly a function of a listener's own production categories. Another finding pointing in this direction is Bigham's (2009) study of the backing of /æ/ in relation to the backing and raising /ɑ/ in young English speakers in Southern Illinois. Bigham tested the hypothesis that /æ/ and /ɑ/ follow a chain shift pattern in which the backward shift of /ɑ/, and its merger with /ɔ/, allow for the backing of /æ/ because /æ/ can now safely occupy a more central position without running the risk of being misperceived as /ɑ/. Bigham found that, indeed, at the level of the community the correlation between the merger of /ɑ/ and /ɔ/ and the backing of /æ/ holds. However, not each individual speaker conforms to the general pattern. There are speakers who produce a clearly backed /æ/ without merging /ɑ/ and /ɔ/. Thus, in these speakers' vowel space, /æ/ and /ɑ/ appear closer than would be expected from the assumption that vowels tend to be most efficiently dispersed in acoustic space. Bigham argues

that these unexpected speakers can safely maintain their system because there is little risk of being misunderstood by others. Their backed /æ/ isn't likely to be misperceived as /ɑ/ as long as enough other speakers do produce /ɑ/ close to or merged with /ɔ/. Crucially, this is only possible if listeners keep close track of the speech of those they interact with and want to be understood by. The exceptional speakers found by Bigham must have some awareness of what variants are more likely and what variants are less likely to be understood, and this awareness appears to be based on their experience with the production of other speakers. Thus, Bigham's results support the view that language users are guided by the variation they are exposed to and not merely the variants which they themselves produce.

The fact that listeners' perception is significantly influenced by their experience of the speech of other speakers is also seen in a number of dialect experience effects reported in the cross-dialect misperception literature which was reviewed at the outset of this section. For example, Labov and Ash (1997) reported that college students showed higher correct recognition rates for non-local variants than high school students. Presumably, what sets high school

students and college students apart is that the latter have experienced more varieties different from their own and therefore have a greater capacity to anticipate and correctly categorize them.

1.2. Sociophonetic knowledge

The results reviewed in the previous section suggest that language users appear to have a latent awareness, or knowledge, of the difference between their own speech and the speech of others. As noted at the beginning of this chapter, I will refer to this knowledge as *sociophonetic knowledge*, or knowledge of how phonetic variation is distributed across speakers in one's speech community.

A central concern of studies concerned with sociophonetic knowledge has been the degree to which it is socially differentiated. That is, how fine-grained is listeners' knowledge of how other speakers and speaker groups differ among each other? This question has been investigated most widely with regard to listeners' accuracy in categorizing different dialects. In a popular research paradigm, listeners hear speech samples of unknown speakers and are asked to

infer some aspect of the speakers' social identity. Examples are regional dialect labeling experiments in which listeners are asked to decide where a given speaker is from. For example, Williams, Garret and Coupland (1999) played samples of different dialects of English spoken in Wales to Welsh teachers and high school students. Overall, the dialects were not identified very robustly, but the teachers showed higher correct labeling rates than the students (see also van Bezooijen & Gooskens 1999 for similar results for Dutch dialects). Clopper & Pisoni 2004b tested Indiana college students' ability to accurately categorize six US regional dialects. They found that while the listeners' general identification accuracy was low, their responses were more accurate than would be expected from chance. Moreover, the listeners displayed an awareness of underlying similarities in the stimuli. Their error patterns accurately reflected similarities between the relevant dialects. The authors also found that those listeners who had previously lived in several other states were more accurate at identifying the dialects from those areas. This again shows that greater exposure to different varieties allows listeners to build up more fine-grained representations of correlations between linguistic and social variation.

Along the same lines, other authors have sought to determine language users' ability to identify ethnicity from speech, especially with regard to African-American and Anglo varieties in the US. In one such experiment, Thomas and Reaser (2004) showed that North Carolina listeners were able to accurately label speech samples from younger African-Americans more often that speech samples from older African-American speakers. They suggest that this is due to the fact that African-American and Anglo speech is fairly similar in the older generations of the relevant speech community, while more recently the two varieties have diverged (see Thomas, Lass & Carpenter 2010 for a more recent summary of the race identification literature). Foreman (2000) studied the identification of African-American ethnicity from prosodic cues. Her findings are particularly interesting because she factored in how close ties participants had with speakers of each variety in terms of their own friendship networks. She found that those listeners who had the most contact with speakers of *both* varieties were the most accurate in their judgments.

Some regional and ethnic labeling studies have also sought to determine what specific linguistic features trigger listeners' decisions. For example, Clopper

& Pisoni (2004a) present an analysis of the stimuli used in Clopper & Pisoni (2004b) and suggest that, for example, the absence of [ɹ] in the word *dark* in one of their stimulus sentences was responsible for the listeners' judgment of the relevant speakers as coming from the East Coast. Other studies used speech synthesis in order to determine a vowel's potential as a social cue. For example, Plichta & Preston (2005) found that listeners were able to distinguish very subtle degrees of monophthongization of the vowel /aɪ/ in the word *guide* when asked to place a speaker on a geographic North-South continuum in the Eastern US. Graff, Labov & Harris (1986) used synthetically manipulated variants of /aʊ/ and /oʊ/ in an experiment where Philadelphia listeners were asked to identify a speaker as either Anglo or African-American. As predicted by the relevant production patterns, more fronted variants gave rise to greater identification rates as Anglo. Going beyond the variables of regional origin and ethnicity, Walker (2007) synthetically created an intrusive [ɹ] following words like *now* in *now it's broken* and also manipulated the release burst of /t/ at the end of intonation units. These two changes led listeners to rate the speakers as younger and lower in social class, respectively.

Research in the dialect labelling tradition has also studied highly localized correlations between linguistic and social variation. For example, Foulkes, Docherty, Khattab & Yaeger-Dror (2010) tested listeners' awareness of the social distribution of variation in voiceless stops in the Tyneside region of Northern England. In the local dialect, male and female speech shows different glottalization patterns for word-medial /p/, /t/, and /k/. Foulkes et al. hypothesized that if local listeners are aware of this correlation between glottalization and gender, they should be able to identify a speaker's gender on the basis of his or her glottalization pattern alone. Non-local listeners, on the other hand, should not be able to do so. To test this, they used recordings of pre-adolescent children because children's voice quality is often gender-ambiguous. As predicted, local listeners' gender judgments were affected by the relevant glottalization patterns while those of control groups of listeners from Southern England and from Arizona were not.

Another study of highly localized variation is Drager's (2009) speech perception experiment at a New Zealand all-girls high school. In her earlier ethnographic research, Drager had found that a major social dimension

separating the girls' friendship groups was the place where they had lunch on school days. She found that this choice had linguistic correlates, notably phonetic aspects of the word *like* used in different grammatical functions. In her experiment, Drager played short excerpts of her interviews with different students back to girls from the same high school. She found that they were able to use some of the relevant phonetic cues to determine where the speaker typically went to have lunch.

1.3 The role of sociophonetic knowledge in speech perception

Overall, the research reviewed in the previous section shows that speakers have extensive, fine-grained knowledge of how sociophonetic variation is distributed in society. However, in each case this conclusion was based on the finding that listeners can reliably access such knowledge when explicitly asked to do so. Listeners were provided with the relevant social categories and then asked to judge whether a given linguistic cue did or did not match it. Thus, the evidence comes from a primarily *social decision,* rather than a primarily *linguistic decision.*

Of course, there are contexts in which making a social decision on the basis of linguistic evidence is precisely what matters, for example in ethnic profiling (Purnell, Idsardi & Baugh 1999). However, this does not reflect most situations in which language is used. Typically, listeners know who they are talking to, and do not need to infer their interlocutor's social identity. Also, from a linguistic perspective it is in the first instance the correct recognition of speech and categories of language that matters, not the recognition of social categories. The studies reviewed above leave open the question whether speech perception may also be informed by knowledge of how speakers from different social groups talk. In this section, I review research which has taken this perspective. I will discuss the relevant studies grouped by the type of task that was used rather than chronologically.

The studies to be reviewed here have all adapted, in one form or another, the matched-guise technique of Lambert, Hogson, Gardner & Fillenbaum (1960). However, unlike in the original matched guise studies of the 1960s and 70s, the variable of the speaker's social identity was not manipulated by recording the same speaker using different languages or language varieties. Rather, different

social guises were created by non-linguistic means, typically by displaying to the listeners a picture or video clip of the ostensible speaker which was then paired with different auditory stimuli across conditions. In this way, the studies to be reviewed here were able to test whether social information by itself, i.e. irrespective of its linguistic manifestation, has an effect on perception.

Strand and Johnson's (1996) work on the perception of the fricatives /s/ and /ʃ/ in male and female speech was the first to show that listeners' linguistic decisions can be influenced by perceived social attributes of a speaker. The authors had previously identified 'untypical' male and female speakers, i.e. speakers whose voices, while being identifiable as male or female, were categorized less rapidly than those of other speakers. In a speech perception experiment, the authors then used these speakers' recordings of the words *sod* and *shod* and systematically manipulated the spectral quality of the initial consonant. They created a synthetic fricative continuum ranging from [s] to [ʃ] and spliced tokens from this continuum onto the original -/ɑd/ sequences. In a perception task, listeners were asked to categorize these tokens as instances of either *sod* and *shod.* To influence the listeners' perception of the speaker's

gender, the auditory stimuli were paired with video clips of male and female speakers producing the relevant words. Strand and Johnson found that the participants were biased toward hearing more tokens as *sod* in the male speaker condition and more tokens as *shod* in the female speaker condition. Thus, the same fricative could be heard as either /s/ or /ʃ/ depending on the perceived gender of the speaker. This shows that the participants had gender-specific assumptions regarding the location of the category boundary between /s/ and /ʃ/ and used this sociophonetic knowledge to categorize words.

A perceived gender effect was also demonstrated by Johnson, Strand and D'Imperio (1999) for the perception of vowel categories adjacent in F1/F2 space. The authors used speech synthesis to create a vowel continuum from [ʌ] to [ʊ] in the words *hud* and *hood* spoken by a gender-ambiguous voice. Again, social information was provided by means of video clips synched to the words. As in Strand and Johnson's (1996) study of fricatives, the gender manipulation induced a category boundary shift. More tokens were categorized as [ʊ] in the female speaker condition and more tokens were categorized as [ʌ] in the male speaker condition. Recently, Glidden & Assmann (2004) replicated this effect.

The above studies of perceived gender effects on speech perception were designed to test theories of speaker normalization. They were not directly concerned with sociophonetic knowledge in the sense discussed here. The finding that listeners have different expectations of male and female speech is not directly related to sociolinguistic variation in a particular speech community. Presumably, it represents knowledge that is widely shared across English dialects and beyond. Its articulatory basis in each case is a universal difference in the size of male and female vocal tracts. A study which went further in the direction of sociophonetic knowledge as pursued here is Drager's (2005, 2011) study of the perception of New Zealand /ɛ/ and /æ/. The background of this study is the prior finding that in recent decades a chain shift has moved both /ɛ/ and /æ/ upwards in F1/F2 space in the speech of New Zealanders of European decent. Today's speakers are differently affected by the shift, with younger speakers showing higher degrees of raising than older speakers. Drager tested whether this asymmetry is reflected in the way the speech of older and younger speakers is perceived by New Zealand listeners. She created synthetic vowel continua from [æ] to [ɛ] in the words *bad* and *bed* as

well as *had* and *head* for both male and female speakers and asked listeners to decide which of the two words they heard. The variable of age was manipulated by displaying to the participants a photograph of a younger or an older person said to be the speaker. As in the studies of gender effects reviewed above, a category boundary shift was observed. Listeners were more likely to categorize vowel tokens as /æ/ in the 'younger speaker' condition than in the 'older speaker' condition, especially the intermediate tokens of the vowel continuum.

A similar age effect was found by Hay, Warren and Drager (2006), also in the context of ongoing sound change in New Zealand. The authors studied the perception of the diphthongs /iə/ and /eə/, already discussed above. Listeners were presented with recorded readings of minimal pairs containing /iə/ and /eə/, such as *here* and *hair*. The speakers who read the stimuli showed varying degrees of vowel merger. In all cases, at least a minimal phonetic distinction was present, i.e. each stimulus could have been correctly identified on the basis of linguistic information alone. The dependent measure was the participants' correct identification rate. The authors manipulated the speaker's perceived age and social class by presenting the auditory stimuli together with photographs of

males and females of different ages and in different attires and surroundings. They found that the listeners' identification accuracy was affected by the photos. Listeners were more likely to confuse /iə/ and /eə/ when the perceived speaker belonged to a group that is, in reality, less likely to maintain the distinction. One interpretation of this result is that, to the degree that potentially disambiguating linguistic information was present, this information was given greater weight in the case of some speakers than others, thus leading to greater error rates for those speakers who are less likely to produce a distinction. Once more, this demonstrates that listeners who are exposed to sociophonetic variation in their community keep track of the likelihood of hearing particular phonetic variants from particular groups of speakers.

Overall, the studies discussed in this section show that sociophonetic knowledge can be used in the opposite direction as in the dialect labeling experiments reviewed earlier. Here, it is ultimately linguistic decisions rather than social decisions that are influenced by speakers' latent assumptions about how phonetic variation is socially distributed. Nevertheless, the tasks used in the experiments reviewed so far leave open a number of questions about the role of

sociophonetic knowledge in other situations. For example, they have in common the overt presentation of the variable studied in the task. The participants may have become aware of the critical manipulation, such as the variable pronunciation of a vowel or consonant by a particular type of speaker. This awareness might have caused the listeners to activate their knowledge of the relevant sociophonetic information to a greater degree than they would otherwise have.

Staum (2008) used a more subtle design in which the nature of the critical variable was less readily identifiable. She used a sentence comprehension task to study the processing of words variably affected by a phonological process which patterns differently in African-American and Anglo dialects of the US. The variable in question was the production or omission of /t/ and /d/ in word-final consonant clusters. In sociolinguistics, this process is known as "t/d-deletion." For example, speakers of African-American English are more likely and speakers of Anglo varieties are less likely to produce words like *mast* as [mæs], rather than [mæst]. In Staum's experiment this variable was introduced covertly. The listeners' task was not to categorize words containing the relevant word-final

stops. Rather, they heard a sentence beginning such as "The [mæs] lasted..." in which the sequence [mæs] is ambiguous between the 'deletion' interpretation ('mast') and the 'non-deletion' interpretation ('mass'). These ambiguous sentence beginnings were followed by a disambiguating continuation, for example "...through the storm" or "... until noon on Sunday." The continuation was designed to disambiguate the word. Staum's response measure was the time it took the participants to determine whether the sentence was semantically well-formed, i.e., to semantically parse the sentence. As in previous studies, the information about the social identity of the speaker was communicated in the form of photographs. However, besides Anglo and African-American photos, the participants also saw photos of speakers of other ethnicities which served as foils. Moreover, the experiment included filler trials with sentences containing no ambiguities due to d/t-deletion. Staum compared the effect which both the disambiguation and the perceived speaker ethnicity had on the listeners' response time. She found that when listeners saw an Anglo speaker, the 'deletion' interpretation took longer to parse than when seeing an African-American speaker. On the other hand, when listeners saw an African-American

speaker, the 'non-deletion' interpretation was parsed more slowly than the 'deletion' interpretation. These results can be interpreted as reflecting the degree of activation which the relevant lexical items, e.g. *mass* and *mast*, received when the ambiguous stimulus, e.g. [mæs], was heard. When seeing an Anglo speaker, who is less likely to have meant *mast*, the word *mass* becomes more strongly activated. This explains the lower response time when the intended interpretation is *mass*. When seeing an African-American speaker, *mast* is more strongly activated because African-American English speakers are less likely to produce a final /t/. This results in quicker responses when the intended interpretation is *mast*. Thus, the results show that listeners are sensitive to the variable likelihood of speakers of these ethnicities to engage in /t,d/-deletion.

The success of Staum's (2008) design shows that sociophonetic effects in speech perception occur even when listeners are unaware of the variable under investigation. Moreover, Staum showed that sociophonetic knowledge can positively affect the speed of processing. However, one shortcoming of this design, in so far as it speaks to speech perception, was that perception was tested in a very indirect way. The actual act of processing the sequences like

[mæs] was not measured directly but indirectly, by means of the speed with which listeners were able to perform a following semantic task. In a study which addressed this point, Koops, Gentry, and Pantos (2008) used an eye-tracking paradigm to measure the online processing of spoken words under the influence of social information about the speaker. Listeners were asked to identify words containing pre-nasal /ɛ/ and /ɪ/, e.g. *dentist* and *dinner*, from different alternatives displayed on the screen. The sociophonetic variable at issue was the merger of the two vowels in this phonological context. The pre-nasal merger of /ɛ/ and /ɪ/ is correlated with age in the dialect studied by Koops et al., the speech of native Anglos in Houston, Texas. Older speakers are less likely to distinguish the two vowels than younger speakers. To test whether listeners assume a merged system to exist in the speech of older but not younger Anglo speakers, the auditory stimuli were paired with pictures of Anglo speakers of different ages. Similar to Hay et al.'s (2006) design, the point of interest was whether listeners who hear potentially merged vowels rely on the linguistic information in the stimulus vowel. In this case, the question was whether listeners who hear, for example, an [ɛ]-like quality will take this to indicate that

the unfolding word actually contains /ɛ/, or whether the phoneme may also be /ɪ/ because the speaker does not distinguish /ɛ/ and /ɪ/. To measure listeners' temporary assumptions as they processed the word, their eye fixations to the target word, e.g. *dentist*, and to a competitor word, e.g. *dinner*, was tracked using a head-mounted eye-tracker. The authors found that, as predicted, listeners spent a greater amount of time fixating the competitor word when listening to an older speaker.

One commonality of the tasks discussed so far in this section is that listeners are asked to identify a word as one of two candidates whose lexical contrast has been, at least potentially, completely neutralized. The listeners task, then, is to resolve global phonological ambiguity in one direction or the other. In the case of a phonetic continuum categorization task like Strand & Johnson's (1996), the linguistic ambiguity takes the form of acoustic tokens which fall on or near the boundary of two adjacent phoneme categories. In the case of vowel merger, as in Hay, Warren & Drager's (2006) study, ambiguity resolution is obviously at issue because vowel merger by definition eliminates linguistic

contrasts. In the case of /t,d/-deletion (Staum 2008), the ambiguity is the result of a phonological process which renders two words homophonous.

The restriction to designs relying on the resolution of global linguistic ambiguity has both practical and theoretical implications. A practical problem is that not all, and probably not even very many, dialect differences can be operationalized in this way. For example, Hay, Warren & Drager (2006) discuss the problem that the number of minimal pairs of words containing /iə/ and /eə/, such as *here* and *hair*, is only barely large enough to create a sufficient number of experimental items. Other socially significant linguistic features do not create lexical contrasts at all. For example, to test whether voice quality such as creaky voice is part of the perceptual representation of some speakers and not others, no method relying on minimal pairs is likely to be found.

But even for variables which produce minimal pairs these designs restrict conclusions to speech perception in the context of global ambiguity resolution. It remains unclear how other types of input are processed in the presence of social information. For example, can the effect also be demonstrated where the critical variable does not result in global ambiguity?

An argument can be made that situations in which linguistic ambiguity activates two lexical candidates nearly of exactly equally are inherently more likely to give rise to social effects on speech perception. Where two competing interpretations are about equally likely from a linguistic perspective, it would be expected that asymmetries in terms of social or other non-linguistic considerations enter the listeners' decision as to how to resolve that ambiguity. But do listeners access social information also where there is no strong lexical competition?

One type of speech perception task used by sociolinguistics that does not rely on lexical ambiguity resolution is a vowel-matching task as employed by Niedzielski (1997, 1999). Niedzielski's hypothesis was that Detroit Anglo listeners' perception of several vowels, including /aʊ/ before voiceless obstruents as in *house*, would be influenced by their assumptions as to whether the speaker is from Michigan or from Canada. Previous language attitude work had indicated that Anglo Detroiters believe that Canadians, but not the Detroiters themselves, produce this vowel with a raised onset, e.g. [ʌʊ]. In Niedzielski's experiment, Anglo Detroiters listened to sentences from a recording

of a Detroit speaker who also produced raised variants. Participants in one condition were told that the speaker was Canadian, while participants in a second condition were told that the speaker was from Detroit. The participants were asked to pay attention to the words containing the relevant vowel and, after having listened to the entire sentence, identify the exact quality of the vowel by picking a variant from a series of synthetically produced vowels presented in isolation. As predicted, the listeners in the first condition were more likely to pick a raised variant than those in the second condition. Thus, it appears that their expectations of the speech of speakers of each nationality are reflected in perceptual biases. Listeners were capable of performing the task consistently in a way that reveals such latent knowledge. The reliability of a vowel matching task in revealing sociophonetic knowledge was corroborated in the recent replication of results similar to Niedzielski's by Hay, Nolan and Drager (2006) and Hay and Drager (2010) in a New Zealand English and Australian English context.

A vowel matching task like Niedzielski's holds the potential to make virtually any phonetic feature amenable to sociophonetic speech perception

experiments. However, it leaves open the question of the exact linguistic processing that leads listeners to their decision. While it is clear that the listeners' motivation for taking into account the social information presented to them is not in order to resolve lexical ambiguity, the nature of a matching task may have a similar effect on their responses as a task based on ambiguity resolution. After all, to correctly match subtle shades of vowel quality also involves dealing with linguistic indeterminacy. Listeners have to decide between competing vowel tokens after a considerable temporal delay while keeping the original vowel quality in memory. This creates a situation in which the effect of sociophonetic expectations may well be stronger than in other situations because listeners are relying more on top-down assumptions than they would otherwise.

In summary, the success of all prior sociophonetic experiments in demonstrating an effect of social information on speech perception may have been at least in part due to the large amount of linguistic indeterminacy created in the tasks. The tasks had in common that the listeners were faced with considerable difficulty in coming to a decision due to linguistic indeterminacy in the stimuli. Thus, an argument can be made that in a situation in which acoustic

phonetic cues point about equally to two alternatives, non-linguistic cues will likely receive greater weight than they would otherwise. The weight of the evidence for one lexical candidate or the other is shifted to social considerations about the perceived speaker. This leaves open the possibility that social information makes a difference only in situations in which listeners are faced with a considerable amount of linguistic indeterminacy, possibly more than in many ordinary processing conditions.

These doubts about the generality of the effect of sociophonetic knowledge on speech perception are echoed in an independent line of research by cognitive psychologists regarding the effects of talker variability on the speed of word recognition (Luce & Lyons 1998; Luce, McLennan & Charles-Luce 2003; McLennan & Luce 2005). The argument comes from long-term repetition priming. In this experimental paradigm listeners are first trained on a set of spoken words. After a temporal delay, they hear another set of words and are, for example, asked to decide whether the words in the second set occurred in the first set. In another version of the task, the participants simply decide whether the words in the second set are English words, i.e. they perform a

lexical decision task. The critical manipulation is that some of the repeated words in the second set are spoken by the same speaker while others are spoken by a different speaker. An effect of talker variability, or indexical variability, can be observed when listeners are more quick to respond to repeated words produced by the same speaker than to repeated words produced by a different speaker. The effect demonstrates that the listeners' decision is facilitated by something other than the lexical information contained in the word because that information is identical. It suggests instead that the listeners' memory representation of the words heard in the first set included, for example, the speaker's voice characteristics.

While indexicality effects of this kind have been demonstrated for some time (Goldinger 1996; Palmeri, Goldinger & Pisoni 1993), they are not found invariably. Luce & Lyons (1998) were unable to replicate the effect in one of their conditions. When their listeners were asked to decide whether the word they heard in the second set was "old" or "new" an indexicality effect emerged. However, when they merely heard the words again as part of a lexical decision task, as explained above, the effect did not emerge. On the basis of Luce &

Lyons's results, Luce, McLennan & Charles-Luce (2003) argue that word recognition is not invariably subject to indexical specificity effects. They suggest that such effects may be absent where processing is rapid, as in a lexical decision task. Where the effect was found, the nature of either the stimuli or the task "may have amplified the effects of voice by either slowing processing or encouraging activation of specific previous memory traces to aid in identification." (203-204) To account for when indexical specificity effects emerge and when they do not, Luce et al. formulate a *time course hypothesis.* According to this hypothesis "the rapidity of responding may mediate the presence or absence of [indexical] specificity effects" (204). McLennan & Luce (2005) tested this hypothesis in another set of long-term repetition priming experiments. The experiments were designed so that the speed with which the listeners processed the stimuli was controlled. As predicted, the authors found that "indexical variability affects participants' perception of spoken words only when processing is relatively slow and effortful." (306).

Luce et al.'s time course hypothesis, according to which indexical specificity effects develop late and may be absent where processing is rapid, can

be readily applied to the results of the sociophonetic experiments described above. Luce et al.'s indexical variability effects correspond in several ways to these effects of sociophonetic knowledge on speech perception. In both cases it is non-linguistic, speaker-specific information that is responsible for the effect, although in one case this information is conveyed acoustically as part of the stimuli themselves (e.g., the speaker's voice quality or articulation rate) and in the other case it may be conveyed separately through another modality (e.g., an image of the speaker). The criticism leveled against the sociophonetic experiments discussed above was that they relied on tasks in which listeners were asked to disambiguate globally ambiguous lexical items. In such tasks responses would be expected to be relatively slow. A processing delay in the case of the most ambiguous stimuli was in fact found in one of the sociophonetic studies cited above. The response times plotted by Drager (2005: 126) show that those tokens which are closest to the perceptual boundary between the vowels /æ/ and /ɛ/ in her experiment were responded to the slowest. Thus, it may be that the effects of social information found in previous sociophonetic experiments of speech perception were due to the fact that, in the words of

McLennan and Luce (2005), processing was "relatively slow and effortful." (306) This leaves open the possibility that where processing is rapid, social information is not accessed in speech perception.

1.4 Hypothesis for this dissertation

How, then, is it possible to determine the role of social information in speech perception independently global ambiguity resolution? One experimental design that meets this criterion is a shadowing task as used by Strand (2000). In Strand's study, listeners heard individual words and were asked to repeat each one as soon as they recognized it. The dependent variable was the participants' response time in repeating the word that was heard. The auditory stimuli were paired with photos of different speakers. Strand's hypothesis was that the social information conveyed by the photo would lead to slower or faster response times if it matched the speech sample in social terms. The photos were headshots of female and male college students, and the auditory stimuli were single words spoken by female and male college students. The pictured students

had been previously determined to be either "typical" or "untypical" of their gender in appearance using the same procedure as for the "typical" and "untypical" male and female voices of Strand & Johnson (1996). The result of Strand's experiment was that seeing a "typical" female photo resulted in quicker word repetition than seeing a "non-typical" female photo. One interpretation of this result is that, at least for female speech, a "typical" female image activates the speech of females more quickly.

Strand's result is interesting as it suggests that listeners' perception is influenced by social information even where there is no lexical ambiguity and, instead, the target words can be identified unambiguously. The listeners merely had to identify the word that was heard out of all words in the English lexicon. The nature of this task is, in this respect, similar to that of a lexical decision task, as used by Luce & Lyons (1998). It is interesting that even under such "easy" processing conditions listeners accessed their sociophonetic knowledge. This finding points to the possibility that sociophonetic knowledge has a much broader role in speech perception than the studies summarized in the previous section were able to demonstrate.

In other respects, however, especially from a sociolinguistic perspective, Strand's findings are inconclusive. Most importantly, it is not clear whether the effect of gender typicality that was found is based on any particular dialect difference between male and female speech. In fact, nothing is known about what features of the voices were responsible for the processing effects. Presumably, pitch differences played a role, but Strand did not systematically manipulate any linguistic feature. Thus, it is difficult to draw inferences from Strand's results to the perception of dialects. Still, the methodology can be easily extended to genuine sociophonetic questions such as those discussed in the previous section.

Strand's finding gives rise to the hypothesis that the processing effect observed for gender typicality is characteristic of the processing of sociophonetic features, i.e. specific linguistic features associated with particular groups of speakers in a particular speech community. If this is the case, prior assumptions about the social category which a speaker is perceived as belonging to will lead to faster recognition of linguistic variants associated with speakers who match that social category. On the other hand, variants which are not associated with

speakers of the relevant social category should be recognized less quickly. I will call the first type of pairing between a speaker and a dialect *congruous* and the second type *incongruous.* A hypothesis can then be formulated as in (1).

(1) Words characterized by sociophonetic congruency are recognized faster than words which are sociophonetically incongruous.

This hypothesis about the processing effects of sociophonetic congruency is independent of whether the alternative values of the sociophonetic variable contained in the word lead to lexical ambiguity. Even where a word is unambiguously identifiable, the speed of processing should be mediated by the variable of sociophonetic congruency.

The purpose of the experiment described in later chapters of this dissertation was to test this hypothesis. The hypothesis was tested in a particular sociolinguistic setting, vowel variation in Houston, Texas. In the next chapter I provide background information on this setting and document several cases in which phonetic variants are demonstrably more likely to occur in the speech of

certain speaker groups than others. These cases of sociophonetic variation are then used to test the hypothesis in (1) in the experiment described in Chapter 3.

Chapter 2

2. Background

In this chapter I discuss sociophonetic variation in Anglo and African-American speakers native to the Houston metropolitan area, specifically variation in the spectral characteristics of vowels. I begin by reviewing prior sociolinguistic and dialectological research on phonetic variation in Texas, with a view to the urban-rural contrast in Anglo dialects (Section 2.1). Next, I present an analysis of selected data from the Houston Urban English Survey (HUES), an ongoing research project with the goal of documenting sociolinguistic variation in the Houston metropolitan area. I first describe the database and methods of acoustic analysis and then illustrate two major axes of variation which emerge from the HUES data. These are age, especially differences between younger and older Anglos, and ethnicity, especially differences between Anglo and African-

American speakers below age 30 (Section 2.2). I then provide a more detailed acoustic analysis of four vowels which constitute particularly clear points of difference (Section 2.3). These four vowels are the ones included in the speech perception experiment described in the later chapters. In Section 2.4 I spell out the specific predictions for speech perception by local listeners from Houston.

2.1. Sociophonetic variation in urban and rural Texas

A major concern of sociolinguistic and dialectological research dealing with language variation in Texas has been the historical development of the Anglo dialect of Texas, including its current trajectory of change. The discussion in this section will be largely restricted to the speech of Anglo Texans because the traditional Texan dialect is most closely associated with Anglo speech in the literature reviewed here. Following Labov, Ash and Boberg (2006), I use the term *Southern*, rather than *Texan* to refer to the relevant dialect features, given that they appear not to be restricted to Texas but, rather, to be pan-Southern features.

One question that has attracted considerable attention is the loss or retention of traditional Southern Anglo dialect features in post-WW2 Texas, and how this is related to the demographic changes which have characterized that time period. Since this dissertation is concerned with phonetic variation in the Houston metropolitan area, one of the largest urban areas in Texas, these findings are directly relevant here. A large research project dealing with urbanization and language change in Texas was conducted in the late 1980s by Bailey and colleagues (Bailey and Bernstein 1989). Their project included a random telephone survey and a survey of high school students. One general result was that some, though not all, traditional Southern Anglo dialect features appear to be receding throughout the state (Bailey 1991, see also Tillery and Bailey 2004). The phonological variables that were investigated include, for example, the monophthongization of /aɪ/, the merger of pre-nasal /ɪ/ and /ɛ/, and the merger of /ɑ/ and /ɔ/ before /ɹ/. In one of the reports of this research, Bailey, Wikle & Sand (1991) distinguish "innovative" variants, i.e. variants which appear to be new to the state, from those which appear to be receding. This determination was made by comparing speakers of different age groups.

Based on the results of their state-wide telephone survey, Bailey et al. argue that the Dallas-Fort Worth metroplex is "the primary focus for the spread of innovations in Texas" (206). For instance, they argue, this area is where one of the innovative changes, the merger of the vowels /ɑ/ and /ɔ/ gained an early foothold.

The importance of large metropolitan areas is specifically discussed by Thomas (1997). While Bailey and colleagues emphasize the linguistic consequences of the social and demographic changes accompanying World War 2, Thomas emphasizes a particular post-war event, the Sunbelt migration (e.g., Abbott 1987). This more recent demographic shift brought large numbers of non-Southern Anglos to the metropolitan centers of the Southern and South-Western states in response to sustained economic growth in the South and the stagnation of traditional industries in the Northern rustbelt states. The onset of the Sunbelt migration in the early 1970s (Tillery and Bailey 2004) seems to have accelerated linguistic changes that were already underway, as well as instigating some new changes. Specifically, as argued by Thomas, this massive demographic change appears to be the reason for the rapid loss of traditional Southern dialect

features among Texas Anglos residing in the large metropolitan centers San Antonio, Dallas, Fort Worth, and Houston. Contact between linguistically Southern and non-Southern speakers in these areas resulted in dialect leveling and the creation of "dialect islands" where a distinctly less Southern, metropolitan Texan dialect is spoken by Anglos, which contrasts markedly from the more traditional dialect which dominates the rest of the state, including both rural areas and smaller cities.

Thomas (1997) acoustically analyzed two vowel variables, /aɪ/ and /eɪ/, because the Southern variants of these variables are among the most often recognized phonological features of traditional Anglo Texan speech (see below for a more detailed discussion of the relevant spectral properties of the variants). Based on data from Bailey and colleagues' random telephone survey and the recordings of high school students from urban and rural areas across the state, Thomas shows a clear synchronic rural-metropolitan split in young Anglo Texans. Moreover, a comparison with older Anglos suggests that this split is a fairly recent development. The urban areas sampled did not include Houston, but some of the interviewees were from suburbs of Dallas.

2.2. The Houston Urban English Survey (HUES)

2.2.1 The HUES word list recordings

The Houston Urban English Survey (HUES, Niedzielski 2006) is an ongoing research project at Rice University with the goal of documenting sociolinguistic variation in the Houston metropolitan area. As an initial step toward that goal, HUES field workers have recorded native Houstonians' readings of a list of words and a reading passage. More recently, spontaneous discourse data have also been recorded. At the time of writing, only the word list data have been fully analyzed. Therefore, the discussion in this chapter will be restricted to the variation found in the word list recordings.

The HUES word list was primarily designed to probe variation in vowel quality. It contains 290 words chosen to be representative of the entire spectrum of English vowels as well as including an emphasis on vowels which are of particular relevance in the context of Southern US dialects. Most of the words

included in the list are monosyllabic. Those which are polysyllabic have lexical stress on the relevant vowel. In most of the words, that vowel appears in a phonological context where the amount of co-articulation with neighboring consonants is minimal: word-initially, word-finally or flanked by oral obstruents. For a small number of vowels, the list also includes words providing pre-nasal and pre-lateral contexts.

The analysis in this chapter is based on 55 word list recordings. They are a subset of the recordings produced to date within the HUES framework. They include recordings of 42 Anglo and 13 African-American Houstonians. The restriction to this subset, and specifically to these two ethnic groups is practical rather than theoretical in nature. The two groups included here are the ones for which the largest samples are available, thus allowing the most accurate estimates to the relevant populations more generally.

The 42 recordings of Anglos include some recordings previously analyzed by HUES researchers with regard to vocalic variation. Eighteen of the Anglo speakers above age 40 formed the basis of Gentry's (2006) analysis of older Anglos, and the 12 youngest Anglos in the sample formed the basis of Pantos'

(2006) analysis of Anglo teenagers. These 30 recordings were the first to be produced in 2006. Since then, additional Anglo as well as African-American speakers have been recorded. Some of these recorded speakers were participants recruited for a related study (Koops, Gentry & Pantos 2008). For purposes of this dissertation, all 55 recordings at issue here were re-analyzed by the author to ensure consistency in measurement technique as well as to introduce additional phonetic measures.

The demographic information available for the 55 speakers discussed here is restricted to their self-reported ethnicity, gender and age. For some of the participants, age at the time of the recording was elicited only in terms of decades, such that they identified their age, for example, as "40s" or "50s". Some potentially relevant demographic variables, such as the speakers' occupation, level of education, or other indicators of socio-economic status, as well as their precise residential history within the Houston metropolitan area were collected for some, but not all participants. Therefore, these social variables cannot be systematically factored into the analysis presented here, and will not be discussed further.

The Anglo group includes 22 male and 20 female speakers. The African-American group includes 6 male and 7 female speakers. The samples differ considerably in terms of the age spectrum covered. The group of African-Americans includes only speakers between the ages of 14 and 20 at the time of the recording. In the Anglo group, the youngest speaker was 14 and the oldest speakers were in their 60s at the time of the recording.

2.2.2. Acoustic measurements

2.2.2.1 Formant duration

The HUES word list readings were analyzed using the acoustic analysis software Praat (Boersma & Weenink 1992-2011). The first step of acoustic measurement was to mark the vowel's beginning and end point. The duration of the vowel was operationally defined as the time period during which both the first and the second formant were clearly identifiable in a wide-band spectrogram as provided in Praat's Editor window. Thus, it would be more accurate to speak of

formant duration than vowel duration. Typically, the first formant was longer in duration than the second formant, having an earlier onset and a later offset. Thus, in the majority of cases, the vowel's beginning and end points correspond to the beginning and end point of the second formant. A major exception are words which end in a high front offglide, e.g. *say*. In such cases, the second formant sometimes extends beyond the first formant, presumably because the frequency range including it is noise-excited, rather than voice-excited, at a breathy-voiced word offset. Another complication are vowels which begin or end in irregularly or widely spaced glottal pulses due to glottalization. In these cases, the pulses were included as part of the vowel because they reflect the spectral properties of the modally voiced parts of the vowel, specifically formant locations.

2.2.2.2 Formant frequencies

Vowel quality was measured acoustically in terms of the center frequencies of the first three formants, as determined by LPC analysis using Praat. The type of

LPC analysis used was the Burg method as implemented in Praat's Editor window. The following strategy was used to arrive at appropriate LPC settings. First, Praat's default LPC settings were applied, whereby five formants are assumed to be present in the frequency range between zero and either 5000 Hertz or 5500 Hertz, for male and female speakers, respectively. Where these settings did not yield a good fit to the energy distribution seen in the spectrogram, LPC poles were added or, less frequently, removed. A "good fit" was defined as a good visual match between the location of the dark bands on the spectrogram and the LPC points provided by Praat. Where this did not resolve unclear cases, a further way of determining a good fit was to check whether adjacent LPC settings yield very similar results. If they did, the LPC model was assumed to be stable and reliable. If they did not, additional adjustments were made.

Having arrived at an appropriate LPC model, the vowel's formant frequencies were measured in two ways. First, measurements of the first three formants were made at a single point of the vowel. How that measurement point

was chosen is described below. Second, for selected vowels formant frequencies were extracted along the entire duration of the first, second, and third formants.

The choice of a point for the single-point measurement followed the methodology of Labov, Ash & Boberg (2006). Whenever possible, the measurement was taken at the vowel's "point of inflection," i.e. at the point where one of the formant contours reverses direction. In the default case, this point was the F1 peak. However, for some vowels additional considerations apply. These are vowels which include a prominent backward or forward gesture in addition to or instead of a prominent downward gesture. For example, /ɔɪ/ in *toy* shows an initial backward gesture, while /u/ in *do* often shows an initial forward gesture. In such cases, the F2 minimum or maximum was taken to be the inflection point.

2.2.2.3 Formant contours

For selected vowels, the entire contour of F1, F2 and F3 was extracted. F1, F2 and F3 values were extracted at 100 equidistant points of the vowel using a

Praat script. Then, an interactive Praat Editor script was used to fit a function to these 100 points. Formant contours were modeled as complex polynomial equations with up to 9 coefficients. The modeling process involved finding an appropriate number of coefficients so that the contour could be fit at an appropriate level of detail. This number was adjusted depending on the complexity of the formant shape. Typically, the number of coefficients used decreased from F1 to F3, given the smaller degree of movement of F3 in most vowels. In the modeling process, points which were clearly mistracked by Praat's LPC algorithm were identified, for example in cases where parts of the vowel were particularly weak in amplitude or noisy. These points were removed prior to fitting the function. One benefit of modeling formant contours, rather than storing the raw formant values, was the reduction in data points to be stored. Moreover, the modeling resulted in a certain degree of smoothing, as in Smoothing Spline ANOVA approaches to formant contours (e.g., Baker 2006, Nycz and De Decker 2006). The relatively large number of up to 9 coefficients was chosen to preserve as much temporal detail as possible. Other authors have found that even lower order polynomials provide reliable fits to formant

contours (e.g., McDougall and Nolan 2007) for the purpose of speaker identification.

In order to make formant frequency measurements comparable across speakers with differently sized vocal tracts, especially male and female speakers, the raw formant frequencies were transformed to a normalized scale. The single-point F1, F2, and F3 measurements were subjected to Lobanov's (1971) vowel normalization method. This method is based on the mean frequency calculated separately for each formant over all measured values. Because the number of vowel tokens in the word list sample was uneven, a mean for each vowel was calculated first, and the grand mean was then calculated over the vowel means. Individual formant frequencies are expressed in terms of standard deviations from the formant mean, or z-scores. These range roughly from -2 to 2.

Lobanov's (1971) normalization algorithm was chosen here because Adank, Smits and van Hout (2004) found it to be the most effective method of abstracting away from physiological effects on formant frequencies while preserving known dialect differences. However, one disadvantage of this method is the level of abstractness of the resulting z-scores. What is lost is the original

formant spacing, i.e. information about the general frequency range in which each formant falls on average. For this reason, a different method was used to normalize the formant contour measurements. This method was the formant-intrinsic version of Nearey's (1978) algorithm, also discussed by Adank et al. (2004). Like Lobanov's method, it is based on the mean frequency of each formant, which was calculated in the same way as above. However, to normalize a particular frequency point, that frequency is multiplied by a speaker-specific scaling factor, which is based on the population mean for each formant. If the speaker sample is composed of both male and female speakers, as is the case here, the resulting normalized scores resemble the formant frequencies of a speaker whose vowel space is of a size intermediate between male and female speakers.

2.2.3. Results: age and ethnicity

The results of the acoustic analysis will be presented in two steps. In the remainder of this section, I discuss overall vowel configurations on the basis of

four illustrative speakers' vowel spaces. The associations between speakers and vowel configurations point to two prominent axes of variation: age-correlated variation within the group of Anglo speakers and ethnic variation across Anglo and African-American speakers. The purpose of this initial discussion is to identify and describe these general trends and to situate them within prior findings on North American dialects. This initial discussion follows the sociophonetic tradition pionieered by Labov, Yeager & Steiner (1972) and developed in subsequent studies (Labov 1994, 2001) whereby phonetic variation in vowel quality is operationalized primarily in terms of each vowel's relative position in a speaker's F1-by-F2 vowel space. Not all observations made here can be substantiated by quantitative analysis. However, in a second step, I will discuss four specific vowels in greater detail. These four vowels' social distribution will be analyzed as exhaustively as possible.

The four speakers discussed below were chosen to be representative of each of the two broad dimensions of variation: age and ethnicity. Each speaker's vowel space is illustrated on the basis of a subset of the vowels included in the HUES word list. Included are the mean F1 and F2 values of all vowels in pre-oral

position. Also included is the mean F1 and F2 value of /u/ occurring before /l/, as in *school*. This vowel context is included because it is of relevance to Southern US dialects, as discussed below, and because in non-Southern dialects it serves to illustrate the high-back corner of the vowel space.

The F1-F2 coordinates in the vowel spaces below represent measurements of the vowel nucleus, or primary inflection point, as described above. As a consequence of reducing the quality of the vowel to this one point, offglide trajectories are not shown. The only exception is the vowel /i/. Here, the F1/F2 coordinate represents a measurement at the glide target, either at the F2 peak or the center of the F2 steady state.

Figure 2.1 and Figure 2.2 show the vowel spaces of two male Anglo Houstonians, a 30-year-old speaker and a speaker in the 50s age group. There are several points of difference between the two speakers' vowel spaces, all associated with Southern phonology. The fact that the Southern features are seen in the older but not in the younger speaker reflects Houstonians' recent reversal of traditional Southern phonology, as discussed above for other metro areas. For instance, a prominent difference is the relative position of the vowels

/eɪ/ and /ɛ/. The two appear "reversed" in the older speaker's vowel space, with a high mid front nucleus of /ɛ/ and a low mid front nucleus of /eɪ/. In the younger speaker's vowel space, by contrast, the two vowels appear in their canonical position.

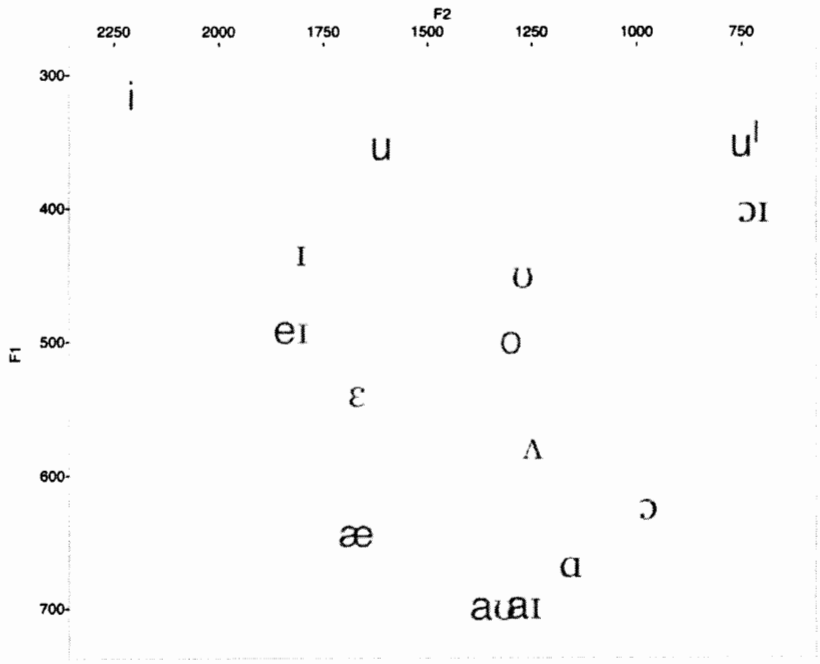

Figure 2.1: Vowel plot of male Anglo Houstonian, age 30

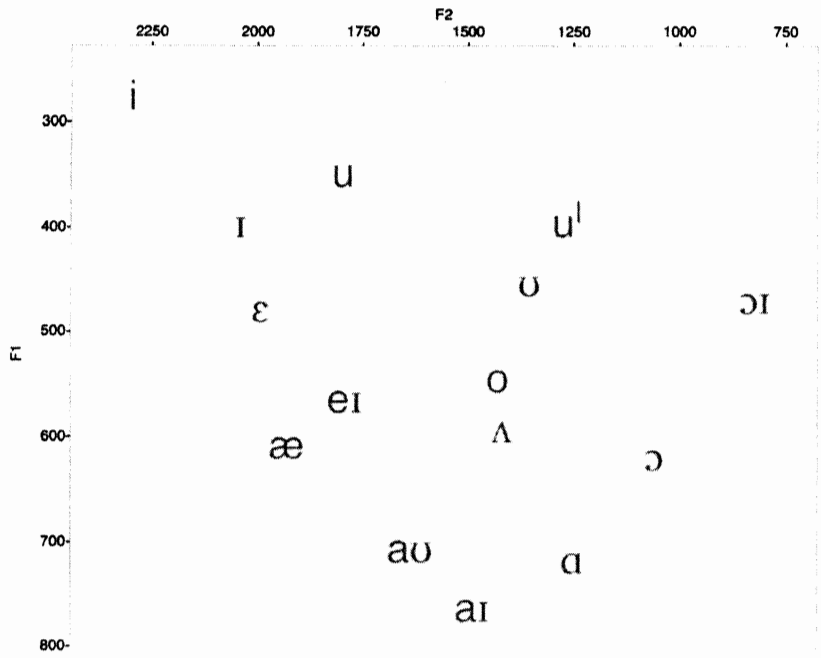

Figure 2.2: Vowel plot of male Anglo Houstonian, age 50s

This "rotation" of /eɪ/ and /ɛ/ in F1-by-F2 space is subsumed by Labov, Ash & Boberg (2006) under the *Southern Vowel Shift* (aka. *Southern Shift*). This sound change, as formulated by the authors, is a sequence of sound changes in three stages. First affected is the vowel /aɪ/, which comes to have an increasingly monophthongal character, ending in a quality close to [a:]. Next affected are the front mid tense and lax vowels. They appear to switch position, as described above. Finally, the high front tense and lax vowels also reverse their position relative to each other in a parallel manner. The speaker shown in Figure 2.2

shows evidence of Labov et al.'s second stage. Therefore, his /eɪ/ and /ɛ/ can be interpreted as Southern and, by extension, traditional Texan dialect features. Another well-known Southern feature in this speaker's vowel space is the fronted position of the vowel /u/ preceding /l/. Labov et al. found this feature exclusively in Southern speakers. Note that in the vowel space of the younger speaker the same vowel appears in a clearly back position. There are other vowel features which distinguish younger and older Anglo Houstonians but are not clearly illustrated in Figure 2.1 and Figure 2.2. These include, for example, the variable merger of the vowels /ɑ/ and /ɔ/, which is widespread among the Anglo teenagers but largely absent in the older Anglos (see also below).

Figure 2.3 and Figure 2.4 show the vowel spaces of two young female Houstonians, a 19-year-old Anglo and a 16-year-old African-American speaker.

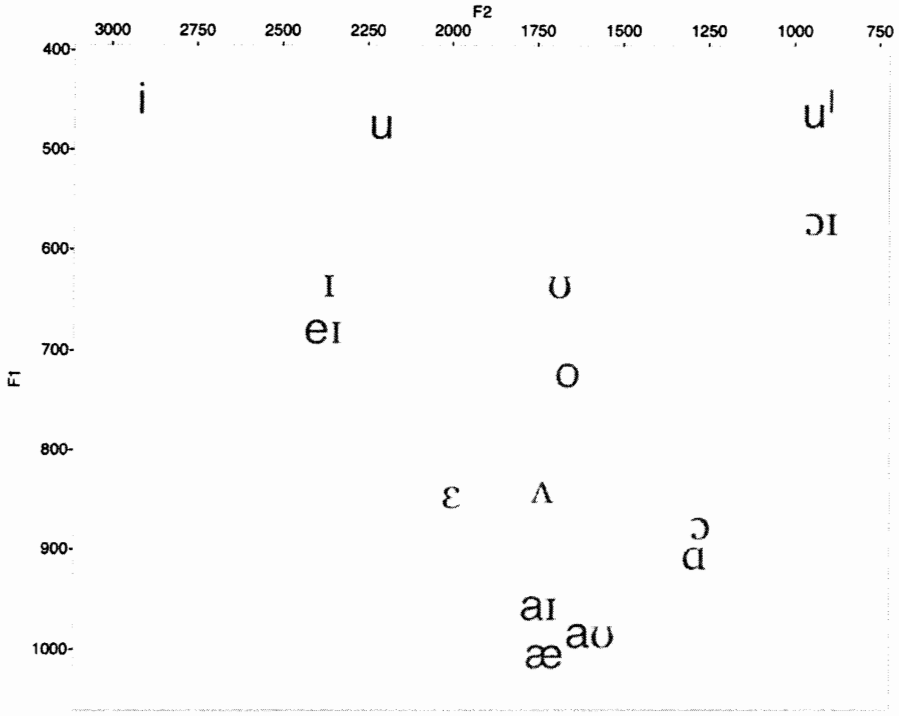

Figure 2.3: Vowel plot of 19-year-old female Anglo Houstonian

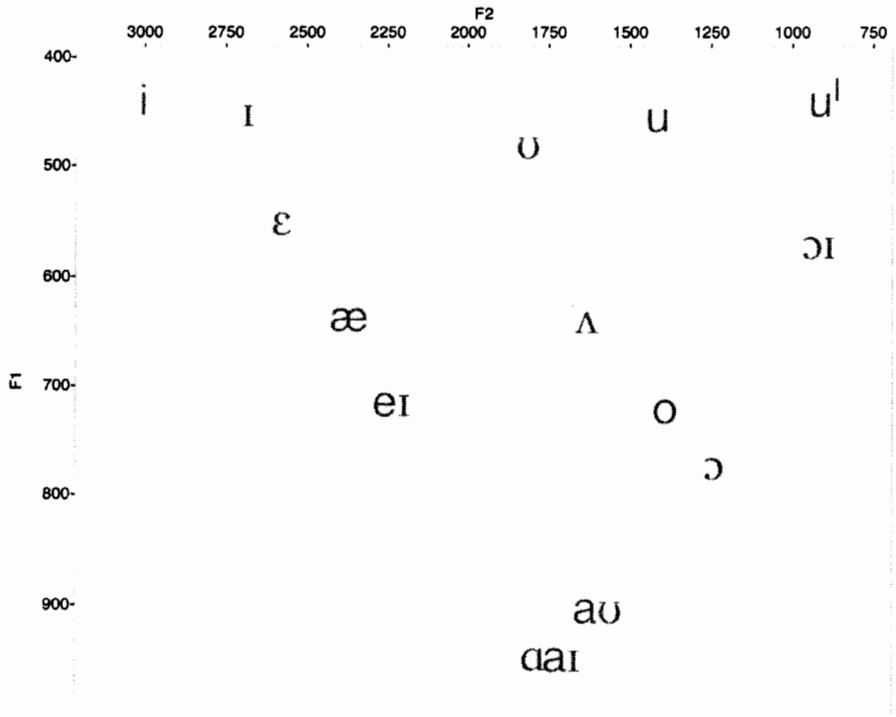

Figure 2.4: Vowel plot of 16-year-old female African-American Houstonian

One clear point of difference between the two vowel configurations is the position of the front lax vowels /æ/, /ɛ/, and /ɪ/. The three appear strongly raised in the case of the African-American speaker but not in the case of the Anglo speaker. Rather, in the Anglo speaker /ɛ/ appears somewhat lowered and /æ/ appears both lowered and retracted. Also showing an opposite pattern is the position of /ɑ/, which is raised and retracted to a position close to [ɔ], and in fact merged with /ɔ/, in the Anglo speaker but centralized to [a] in the African-

American speaker. As noted above, for most of the teenage Anglos in the present sample, the vowels /ɑ/ and /ɔ/ are merged in production. Another clear point of difference is the position of /ʊ/ and /ʌ/. In the case of the African-American speaker, both of them appear raised relative to their position in the Anglo speaker's vowel space. Note that /ʊ/ is well above /ɔɪ/ in her speech, but considerably lower than /ɔɪ/ in the speech of the Anglo speaker. Similarly, /ʌ/ is considerably higher than /o/ in her speech, but lower than /o/ in the speech of the Anglo speaker. Finally, the position of the vowels /u/ and /o/ is less front in the speech of the African-American speaker.

The African-American speaker's /ɪ, ɛ, æ, ɑ/ configuration is consistent with Thomas' (2007) hypothesized *African American Shift.* In this sound change, which Thomas tentatively analyzes as a chain shift, all three front lax vowels are raising and /ɑ/ moves to a low central position. Parts of this shift, especially the raising of /æ/ and the fronting of /ɑ/ have been noted widely in African-American varieties across the US (Yaeger-Dror and Thomas 2010).

2.3 Further analysis of selected vowels

The dialect differences emerging from the HUES word list data include a wealth of contrasts. In the rest of this chapter, I further discuss four vowel variables which constitute particularly clear points of difference, across age, in one case, and across ethnicity, in the other case. I have chosen two variables each from the dimension of Southern vs. non-Southern speech (here, correlated with speaker age) and the dimension of Anglo vs. African-American speech, for younger speakers. This analysis will provide more conclusive evidence showing that each vowel is indeed socially distributed in the way suggested by the illustrative vowel spaces discussed above.

The quantitative analysis below consists of a series of linear mixed-effects regression models, whereby the F1 and F2 values of all relevant words in the HUES data are modeled on the basis of the social information available for the speakers: their age, gender, and ethnicity. The model building techniques follow the guidelines for mixed-effects regression modeling in Baayen (2008) using the *lmer* function of the *lme4* package (Bates & Sarkar 2007) in the statistical

software *R* (R Core Development Team 2010). Mixed effects regression models have recently gained popularity in variationist sociolinguistics (e.g., Johnson 2009) because of their ability to compensate for the potentially misleading effects of grouping variables such as the particular speaker who produced a speech sound and the particular word containing it. Sociolinguistic data sets typically consist of many observations taken from a relatively small number of speakers. Including 'speaker' as a random effect prevents between-speaker variables, such as speaker age or speaker gender, from being over-estimated. Similarly, including 'word' as a random effect prevents the effects of individual items to exert undue influence. In the regression models presented below, the variables 'speaker' and 'word' were entered as random effects.

As discussed by Baayen (2008), one problem in evaluating the results of mixed effects models produced by the functions in the *lme4* package is that, unlike regular regression functions, they do not provide p-values for t- and F-tests. The reason is that it is unclear how to calculate the relevant degrees of freedom in mixed models. Following Baayen, I have therefore used the *pvals.fnc* function of the *languageR* package, which estimates *p*-values and confidence

intervals for the *t*-statistic by means of Markov chain Monte Carlo sampling. The significance level was defined as 0.05.

2.3.1. /eɪ/ and /ɛ/ in Anglo speakers

The age-correlated variation in the quality of /eɪ/ and /ɛ/ likely constitutes a fairly unique Houston feature as it is directly related to the city's demographic history, specifically the rapid loss of traditional Southern phonological features over the past half century. While similar configurations would be expected in other large metropolitan areas in Texas and elsewhere in the South, this pattern would not be expected outside of the South, where Southern vowel variants may not occur at all, or in the rural South, where Southern vowel variants are not receding but may in fact be increasing (Thomas 1997). Figure 2.5 is a plot of the mean normalized F1 and F2 values of /eɪ/ of all 42 Anglo speakers. The mean values shown are based on 22 words containing stressed /eɪ/: *date, bait, hate, pace, fake, cake, take, bade, paid, game, baker, hey, hay, day, stay, jay, say, gay, Kay, okay, bay* and *pay*. These are all the words containing /eɪ/ in the HUES

word list except for *away*, which was excluded here because of the particularly strong influence of the preceding /w/ on the formant frequencies of the vowel. Visual inspection of Figure 2.5 suggests a clear age effect, but no obvious gender effect.

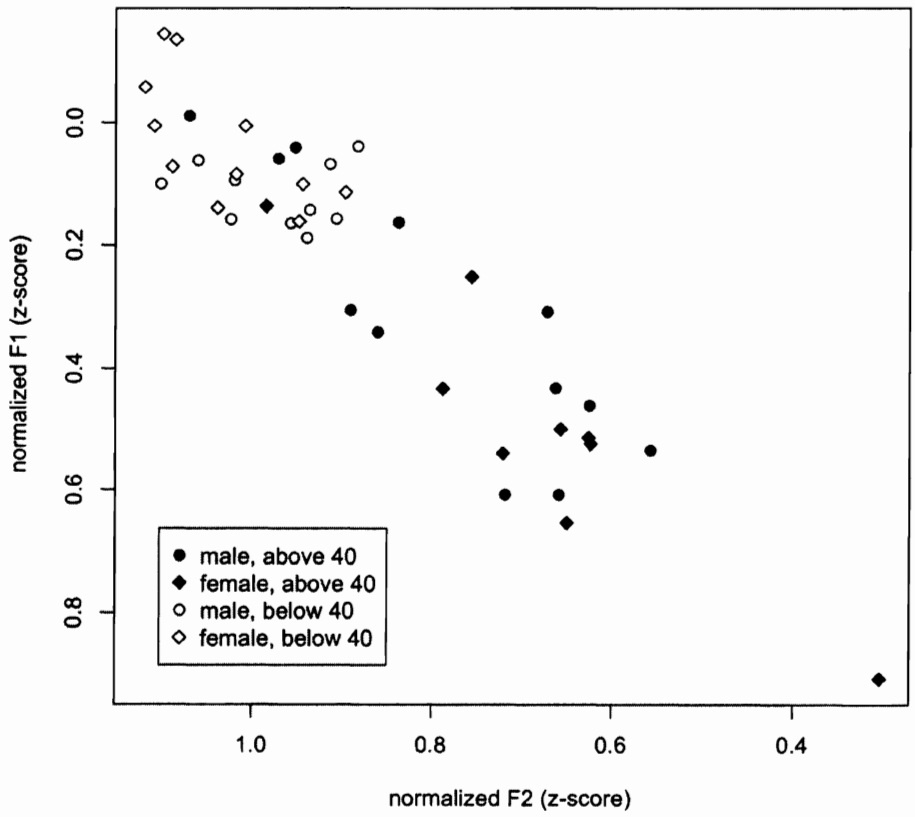

Figure 2.5: Mean normalized F1 and F2 of /eɪ/ for all 42 Anglo speakers

To test this and other possible correlations, separate linear mixed effects regression analyses were conducted for F1 and F2 with the speakers' age as a continuous fixed effect and the speakers' gender as binary fixed effect. A possible interaction between age and gender was also tested. The significant fixed effects of each model are shown in Table 2.1.

Predictors of F1	Estimate	Std. error	*t*-value
Age	0.013236	0.001627	6.363***
Predictors of F2	**Estimate**	**Std. error**	***t*-value**
Age	-0.008408	0.001245	-6.756***

Table 2.1: Fixed effects in regression models fit to F1 and F2 of /eɪ/. Symbols following the *t*-value indicate the associated *p*-value: '***' $p<0.001$, '**' $p<0.01$, '*' $p<0.05$, '.' $p<0.1$

As seen in Table 2.1, both F1 and F2 are strongly predicted by age, but not by speaker gender. The older Anglos' /eɪ/ in the HUES sample has a higher F1 and a lower F2, corresponding to a lower and more retracted articulation, more like [ɛɪ] or, in the case of one speaker, [æɪ].

Figure 2.6 shows the normalized mean F1 and F2 values of /ɛ/ for all 42 Anglo speakers. The means shown in Figure 2.6 are based on all 10 words in the HUES word list containing /ɛ/: *jet, set, pet, bet, kept, dead, Fed, Ted,* and *ever.* Visual inspection indicates a clear age effect, in the opposite direction as in the case /eɪ/. This reflects the "rotation" of these vowels in the vowel space of linguistically Southern speakers, as discussed above.

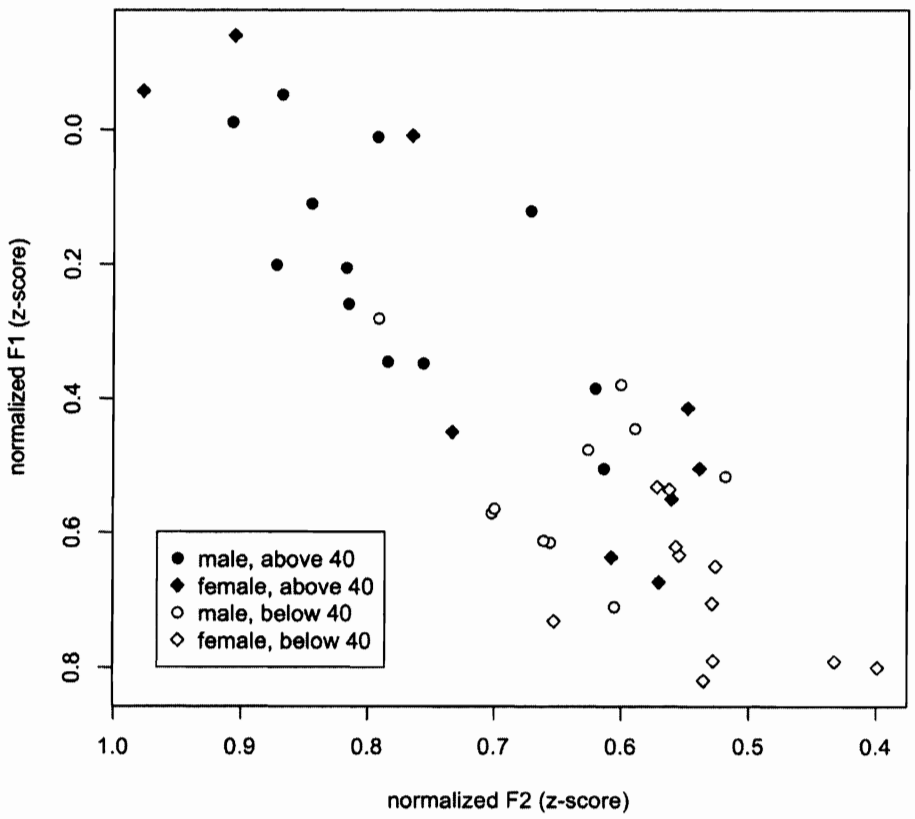

Figure 2.6: Mean normalized F1 and F2 of /ɛ/ for all 42 Anglo speakers

To test the age effect for statistical significance, separate linear mixed effects regression models were fit to the normalized F1 and F2 values of all /ɛ/ tokens represented in Figure 2.6, with speaker age as a continuous predictor variable and speaker gender as a binary predictor variable. A possible interaction between age and gender was also tested. The two models are summarized in Table 2.2.

Predictors of F1	Estimate	Std. error	*t*-value
Age	-0.010369	0.001932	-5.367***
Predictors of F2	Estimate	Std. error	*t*-value
Age	0.004985	0.001063	4.692***

Table 2.2: Fixed effects in regression models fit to F1 and F2 of /ɛ/. Symbols following the *t*-value indicate the associated *p*-value: '***' $p<0.001$, '**' $p<0.01$, '*' $p<0.05$, '.' $p<0.1$

As seen in Table 2.2, there is a main effect of age. The older speakers' /ɛ/ nucleus has a lower F1 and a higher F2, or, in articulatory terms, a higher and further advanced /ɛ/, more like [e] than [ɛ].

2.3.2 /ɑ/ and /ʌ/ in African-American and Anglo speakers

As discussed above, a particularly clear dialect contrast between teenage African-American and Anglo speakers in Houston is the opposite patterning of the front lax vowels /ɪ, ɛ, æ/ and the low back /ɑ/. For the youngest speakers /ɪ, ɛ, æ/ appear to be strongly raised and /ɑ/ appears clearly centralized. In the corresponding Anglo vowel space, especially for female speakers, /æ/ appears lowered and retracted, while /ɑ/ is retracted and raised. While neither of these vowel patterns is completely unique to Houston, their combination can be viewed as somewhat unique. As discussed above, the African American Vowel shift is to some extent attested in other areas, but clearly not to the same extent and with the same regularity as in the data discussed here. Also, strong raising of /ʌ/ has rarely been reported in the literature on varieties of African-American English (see for example the papers in Yaeger-Dror and Thomas 2010).

Figure 2.7 shows the mean F1 and F2 of /ɑ/ of all African-American and Anglo speakers below age 30 in the sample.

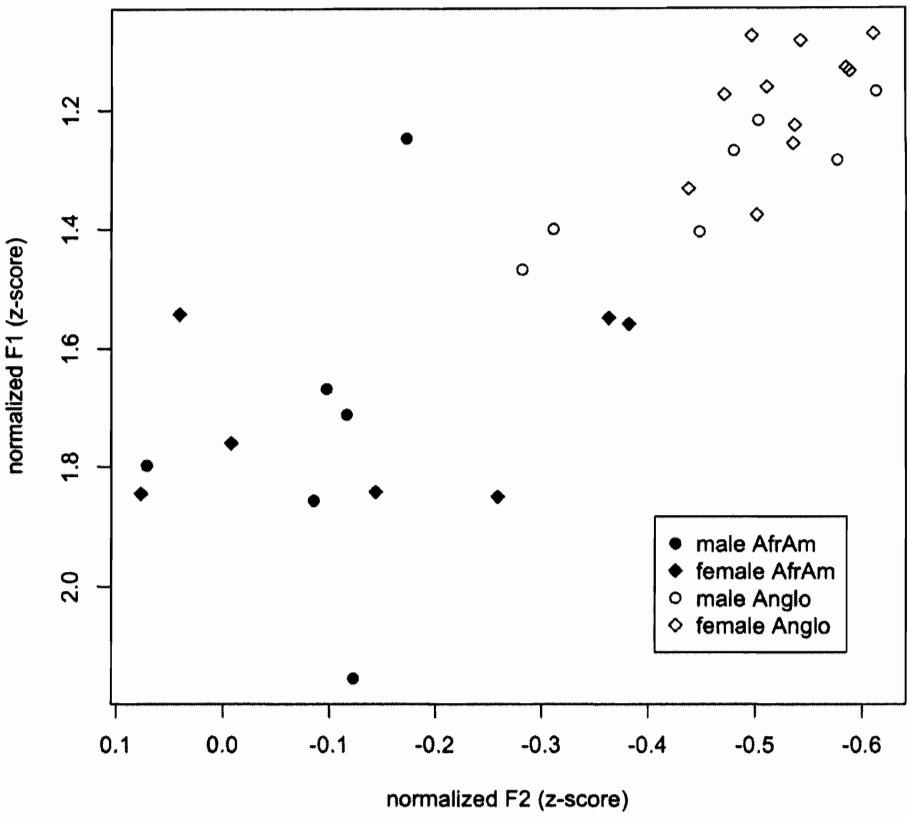

Figure 2.7: Mean normalized F1 and F2 of /ɑ/ for all African-American and Anglo speakers below age 30

The set of words in the HUES word list representing the vowel /ɑ/ was relatively small, containing only 7 items: *box, copy, cot, hockey, pot, got* and *not.* The grammatical function words *got* and *not* each showed some word-specific effects, both having a noticeably lower and fronter target. However, this effect was, impressionistically, very consistent. That is, it did not vary noticeably with the

dialect of the speaker. Therefore, and given the small number of words containing /ɑ/, they were not excluded from the analysis.

As can be seen in Figure 2.7, there is a clear effect of speaker ethnicity with almost no overlap in F1 or F2 between Anglo and African-American speakers. To test this and other possible correlations between formant frequencies and social variables, separate regression models were fit to the F1 and F2 values, with the speakers' ethnicity, age, and gender as fixed effects. All pairwise interactions were also tested. The results are summarized in Table 2.3.

Predictors of F1	Estimate	Std. error	*t*-value
Ethnicity Anglo	-1.56216	0.40849	-3.824***
Age	-0.04770	0.01994	-2.393*
Ethnicity Anglo : Age	0.06293	0.02354	2.674**
Predictors of F2	**Estimate**	**Std. error**	***t*-value**
Ethnicity Anglo	-1.17319	0.28307	-4.145***
Age	-0.02507	0.01381	-1.815.
Ethnicity Anglo : Age	0.04535	0.01631	2.780**

Table 2.3: Fixed effects in regression models fit to F1 and F2 of /ɑ/. Symbols following the *t*-value indicate the associated *p*-value: '***' $p<0.001$, '**' $p<0.01$, '*' $p<0.05$, '.' $p<0.1$

Table 2.3 shows a main effect of ethnicity, in both F1 and F2, in the expected direction. The African-American speakers' /ɑ/ has both a higher F1 and a higher F2, corresponding to a lower and more central articulation. In addition, there is a significant but considerably weaker interaction between ethnicity and age for both F1 and F2. For F1, but not for F2, there is also a main effect of age. The main effect of age was retained in the F2 model because of the interaction with ethnicity. The directionality of the interactions is such that with increasing age each group shows a reversal of the general contrast. In other words, with greater age, the Anglos' and the African-Americans' /ɑ/ becomes more similar. However, at least for F1, the mutual convergence is not equally strong for both groups. The direction of the main effect of age is such that it largely offsets the reversal for the Anglos seen in the interaction of age and ethnicity. Overall, then, /ɑ/ is clearly correlated with ethnicity, but the contrast is strongest for the youngest speakers in the sample and becomes slightly attenuated with age primarily because the African-American speakers closer to their 20s show a slightly less fronted /ɑ/.

Another vowel representing a clear point of variation across ethnicity is /ʌ/ because of the low F1 of many of the African American speakers in the HUES sample. Figure 2.8 shows the F1 and F2 means of /ʌ/ for all speakers below age 30.

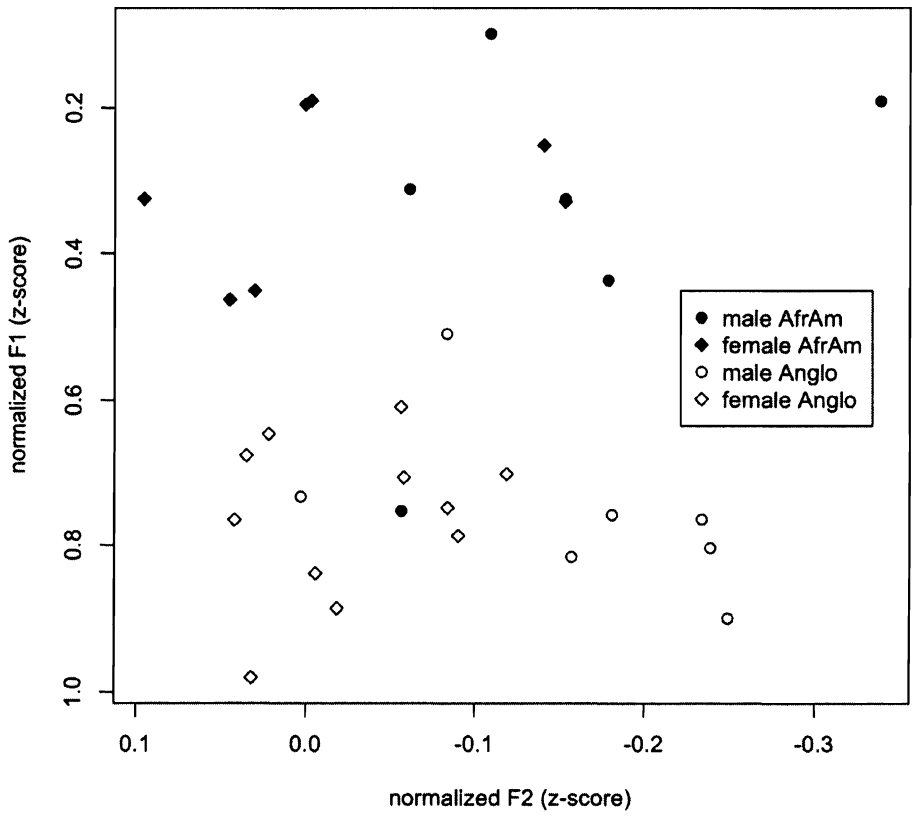

Figure 2.8: Mean normalized F1 and F2 of /ʌ/ for all African-American and Anglo speakers below age 30

The words in the HUES list containing /ʌ/ are *cut, puck, putt, but, dud, bud, tug, pug, nothing, mother.* The grammatical function words *but* and *nothing* each showed a slightly word-specific distribution, *but* being higher in F1 and *nothing* being more advanced in F2. However, there was no clear indication that this effect differed systematically across speakers. Therefore, to boost the relatively small number of /ʌ/ words in the sample, these words were not excluded.

Again, separate mixed effects regression models were fit to the F1 and F2 data. The models are shown in Table 2.4.

Predictors of F1	Estimate	Std. error	*t*-value
Ethnicity Anglo	1.11470	0.31957	3.488***
Age	0.04413	0.01556	2.836**
Ethnicity Anglo : Age	-0.04082	0.01840	-2.219*
Predictors of F2	**Estimate**	**Std. error**	***t*-value**
Gender male	-0.125676	0.026468	-4.748***
Age	0.0274	0.009213	2.970**
Ethnicity Anglo	0.375822	0.191194	1.966.
Ethnicity Anglo : Age	-0.0242	0.010986	-2.203*

Table 2.4: Fixed effects in regression models fit to F1 and F2 of /ʌ/. Symbols following the *t*-value indicate the associated *p*-value: '***' $p<0.001$, '**' $p<0.01$, '*' $p<0.05$, '.' $p<0.1$

In the F1 dimension, the fixed effects closely resemble those in the models for /ɑ/. First, there is a strong main effect of ethnicity. The African-American speakers' /ʌ/ target is higher, closer in height to [ʊ]. In addition, there is both a main effect of age and an interaction of age and ethnicity which together show that the African-American /ʌ/ approaches that of the Anglos with age. The Anglo /ʌ/, by contrast, is stable with age because for the Anglo speakers the effect of the interaction between age and ethnicity is fully offset by the main effect of age. In the F2 dimension, there is no significant main effect of ethnicity. Ethnicity was retained as a variable in the model only because it interacts with age. The main social correlate of F2 is gender. The mean male /ʌ/ appears slightly back of that of the female speakers. There is also a small positive effect of age on F2, which is, however, offset by an interaction between age and ethnicity in the case of the Anglos, resulting in a negative effect of age on F2 in the African-American speakers. That is, with greater age the F2 of the African-Americans' /ʌ/ becomes further fronted.

2.4 Summary and predictions for speech perception

In this chapter I presented a close analysis of four cases of sociophonetic variation in the Houston metropolitan area: the variable production of the vowels /eɪ/, /ɛ/, /ɑ/ and /ʌ/. The first two cases pertain to the speech of Anglos of different age groups and are indicative of the reversal of traditional Anglo dialect features by metropolitan Texans in recent generations (Thomas 1997). Many Anglo speakers above age 40 produce the vowels /eɪ/ and /ɛ/ in what has been described as a "rotated" configuration (Labov, Ash & Boberg 2006). Younger Anglos, by contrast, produce more canonical, non-Southern variants of the two vowels. The second point of sociophonetic variation is seen in comparing the speech of Anglo and African-American Houstonians. Here, the vowels /ɑ/ and /ʌ/ show particularly strong and consistent differences. Judging by the HUES sample, the vowel /ʌ/ is frequently raised in the speech of young African-Americans but consistently not raised in the speech of young Anglos. The vowel /ɑ/ is raised and backed in the speech of Anglos, but lowered and centralized in the speech of African-Americans.

In Chapter 1, the term *sociophonetic congruency* was used to describe instances of the use of a sociophonetic variant by a speaker belonging to a social group which is, in the experience of a listener, likely to occur. Conversely, unlikely pairings of social and phonetic variation were called *incongruous*. For example, applied to the Houston speech community where a raised /ʌ/ is statistically associated with the speech of one social group, African-Americans, but not with the speech of another group, Anglos, the use of a raised /ʌ/ by an African-American speaker can be described as congruous, while the use of a raised /ʌ/ by an Anglo speaker can be described as incongruous.

The research reviewed in Chapter 1 shows that speech perception can be informed by sociophonetic knowledge. Given that sociophonetic congruency forms part of a listener's larger body of sociophonetic knowledge, congruency should also hold the potential to influence speech perception. Specifically, as discussed in connection with Strand's (2000) work on the perceptual effects of gender typicality, there is some evidence to suggest that sociophonetic congruency can have a facilitative effect on word recognition. Word recognition appears to be faster in cases of congruency than in cases which lack congruency.

In the experiment reported in following chapters, the four vowels isolated in this chapter were used as test cases to investigate the effect of sociophonetic congruency on the speed of word recognition. In accordance with the hypothesis that sociophonetic congruity mediates the speed of word recognition, several predictions were tested regarding the way native Houstonians perceive these four vowels when spoken by other Houstonians. The predicted effect consists in an interaction between vowel quality and the perceived social characteristics of a speaker. The perception of the vowels /eɪ/ and /ɛ/ in Houston was predicted to show an interaction between the perceived age of the speaker and the degree to which each vowel exhibits the phonetic properties associated with the Southern Vowel Shift. That is, a lowered variant of /eɪ/ and a raised variant of /ɛ/ should each be recognized more quickly if the perceived speaker is an older Anglo than if the speaker is perceived as a younger Anglo. On the other hand, canonical variants of /eɪ/ and /ɛ/ should be recognized more quickly when the speaker is perceived as younger and more slowly when the speaker is perceived as older. The perception of /ɑ/ and /ʌ/ should show an interaction effect between the quality of the vowel and the speaker's ethnicity. If the speaker is a

young African-American, a raised variant of /ʌ/ and a fronted variant of /ɑ/ should each be recognized more quickly than the same variants spoken by a young Anglo speaker. On the other hand, in the speech of a young Anglo speaker a non-raised /ʌ/ and a backed /ɑ/ should have a processing advantage relative to the same variants produced by a young African-American speaker.

According to the research hypothesis formulated in Section 1.4, the speed of word recognition is mediated by sociophonetic congruency even under easy processing conditions. Specifically, it was predicted that sociophonetic information should have an effect even where a listener does not have to disambiguate a globally ambiguous stimulus. The two pairs of vowel variables discussed in this chapter, /eɪ/ and /ɛ/ as well as /ɑ/ and /ʌ/, lend themselves to test this prediction. They are appropriate test cases because in neither case can one of member of the pair be easily confused with the other. The phonetic range of variation spanned by each does not include variants which could be mistaken as a variant of the other. In the case of /ɑ/ and /ʌ/, even the lowest /ɑ/ variants don't overlap with any variants of /ʌ/ as can be seen in the vowel plots of the advanced speakers in Figures 2.1 and Figure 2.2. In the case of /eɪ/ and /ɛ/, the

internal trajectory of the vowels differs dramatically in that /eɪ/ shows a front upglide in all variants, while /ɛ/ does not. Thus, the phonetic properties of these two sets of vowels differ significantly from, for example, the New Zealand English vowels /æ/ and /ɛ/ studied by Drager (2005, 2011). The acoustic space occupied by /æ/ and /ɛ/ in F1-by-F2 space is directly adjacent, so that a strongly raised /æ/ can be easily interpreted as a conservative /ɛ/.

Chapter 3

3. Methodology

An experiment was designed to test the predictions formulated at the end of Chapter 2. The aim of the experiment was to determine whether Houston listeners show a processing difference in their phonetic perception of Southern and non-Southern vowel variants, in one case, and Anglo and African-American vowel variants, in the other case, depending on the perceived social identity of the speaker.

3.1 Matched-guise design

Following other sociophonetic speech perception experiments (see Chapter 1), the current study made use of a variant of the matched-guise technique in which

auditory stimulus words were presented to the participants as having been uttered by different speakers so as to determine whether the variable construal of the speakers' social identity influences speech processing. The speakers' social identities were communicated to the participants primarily by displaying different photographs on the screen. Furthermore, the instructions given to the participants before and during the task were designed to reinforce in the minds of the participants the impression that the speakers were real Houstonians with the relevant social characteristics. These aspects of the experiment are discussed below.

Participants heard four voices. There were two male and two female voices. I will refer to them as *male-1*, *male-2*, *female-1*, and *female-2*. Each voice was consistently paired with one of four color photographs of the ostensible speaker. The photographs were headshots of two male Anglos, one younger and one older, and two females, one young Anglo and one young African-American. The photographs were chosen to match as closely as possible the relevant age and ethnicity characteristics (see below for the fictitious ages). Despite being fictitious, I will continue to refer to these voice-picture pairing as a "speakers."

The matched guise design required there to be two separate groups of participants, which I will refer to as *Group A* and *Group B*. All participants, regardless of which group they were in, heard exactly the same auditory stimuli in the same order. The only difference between the groups was the matching of auditory and visual stimuli, as shown in Table 3.1.

Voice	Group A Photo	Group B Photo
male-1	older Anglo	younger Anglo
male-2	younger Anglo	older Anglo
female-1	young AfrAm	young Anglo
female-2	young Anglo	young AfrAm

Table 3.1: Voice-photo pairing in the matched-guise design

3.2. General procedure

In order to precisely control the spectral characteristics of the dialectal variants heard by the participants speech synthesis was used to prepare the auditory stimuli. The synthesis process is discussed in detail in the later sections of this

chapter. The technical challenges and time demands of highly realistic speech synthesis allowed only a limited number of stimuli to be produced. As a result, the number of words which the participants heard the speakers produce was rather small. Each speaker was heard saying only four different words. This very small stimulus set influenced the choice of instrument to determine the speed of word recognition. It was decided that such a small set size would defeat the potential advantage of a shadowing task as used by Strand (2000). Among the advantages of a shadowing task is that listeners are not provided with explicit response alternatives. In principle, the response set is the entire lexicon. This makes a shadowing task a good simulation of word recognition in real-life contexts. However, if the set of target items is very small, as in the current study, such that the same items keep reappearing, listeners will likely infer the size of the response set after a few trials. This effectively defeats the advantage of a shadowing task.

Given these considerations, the participants in the current study performed a task that was easier to implement than a shadowing task. They performed a two-alternative forced choice word identification task in which they

saw two response alternatives on a screen in front of them. They responded by pressing one of two buttons which were mapped to the words. The task was administered on an Apple Macintosh laptop computer running the stimulus presentation software SuperLab 4 (Cedrus Corporation, San Pedro, CA). The auditory stimuli were heard through Sennheiser HD 280 pro headphones. The computer sound volume was set to a comfortable listening level that was held constant across participants. The participants responded by pressing one of two adjacent buttons on a Cedrus RB-830 USB button box. They were instructed to identify the word they heard as quickly as possible by pressing either the left or the right of the two buttons.

Throughout the duration of each trial, a color photograph of the speaker, 1.75 by 1.75 inches in size, was displayed in the center of the screen in front of a light gray background. One second after the beginning of a trial, a sound file containing a single word started to play. The trial ended with the participants' response. If the response occurred prior to the end of the sound file the trial ended after the file had played to the end. If no response occurred within

3000 ms of the onset of the sound file, the trial ended automatically and the next trial started.

There were 16 blocks of trials. Each block contained 24 trials. This amounts to a total number of 384 trials. Participants were able to pause between blocks. Throughout each block, the two response alternatives and their mapping to the left or right button remained constant, but the mapping was provided only before the start of the first trial. The first and the last eight blocks were identical in terms of the auditory and visual stimuli, but the mapping of the response alternatives to the buttons was reversed, as shown in Table 3.2.

		Response alternatives				Response alternatives	
Block	Voice	Left	Right	Block	Voice	Left	Right
1	male-1	*bay*	*bed*	9	male-1	*bed*	*bay*
2	female-1	*duck*	*dock*	10	female-1	*dock*	*duck*
3	male-2	*day*	*dead*	11	male-2	*dead*	*day*
4	female-2	*stuck*	*stock*	12	female-2	*stock*	*stuck*
5	male-2	*bay*	*bed*	13	male-2	*bed*	*bay*
6	female-2	*duck*	*dock*	14	female-2	*dock*	*duck*
7	male-1	*day*	*dead*	15	male-1	*dead*	*day*
8	female-1	*stuck*	*stock*	16	female-1	*stock*	*stuck*

Table 3.2: Voices, response alternatives, and mapping of response alternatives to left or right button in each experimental block

Before the start of the first trial in each block the participants saw the upcoming speaker, the response alternatives, and the mapping of the response alternatives to the left and right buttons, as shown in Figure 3.1.

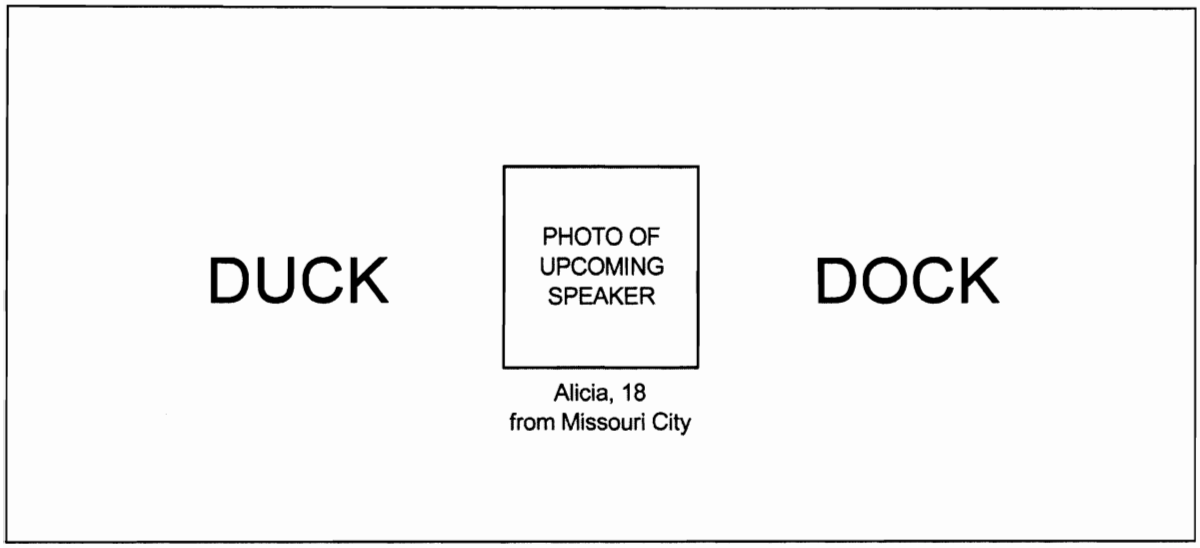

Figure 3.1: Sample visual display prior to the beginning of a block

As can be seen in Figure 3.1, in addition to a photo image, each speaker was also identified by a first name, an age, and a regional label. Neither this additional information nor the response alternatives were seen during the trials. These additional social cues are discussed below.

The 16 experimental blocks were preceded by a practice block containing 12 trials. The practice trials were identical in format to the experimental trials but featured a different speaker and different lexical items. The practice speaker was a fictitious Asian-American speaker of age 19. This social identity was chosen because none of the experimental blocks involved the speech of either

Asian speakers or teenage male speakers. The lexical alternatives in the practice trials were the words *cap* and *window*. These were chosen because the vowels included in these words did not occur in any of the experimental lexical items. The auditory stimuli were recorded readings of the words *cap* and *window* by a college-age Asian-American speaker who had read the HUES word list.

At the start of the experiment, prior to the practice and experimental blocks, several kinds of demographic information were collected from the participants through the same computer interface that they used to perform the experimental task. The participants were first asked to enter the year in which they were born and to specify their gender and their ethnicity/race in an open response format. Next, they were asked whether English was their native language, which was defined as a language they learned to speak at about age 2 and have continuously spoken since then. They were also asked whether they spoke any other languages natively, and if so, which ones. The next question asked whether the participants had grown up, between age six and 18, in the Houston metropolitan area. This was defined as the city of Houston and all surrounding cities and suburbs within about 30 minutes that are urban or

suburban, but not rural, in character. In addition, they were asked whether they had lived in the Houston metropolitan area continuously since age 18 and, if not, for how many years they had lived elsewhere. Finally, they were asked whether they were currently experiencing any speech or hearing problems.

The speech perception task of the experiment was introduced by explaining to the participants the following about the background of the study.

> For the past four years, researchers at Rice University and the University of Houston-Downtown have interviewed Houstonians from all over the Houston metropolitan area. You will hear some of their voices today. The people you will hear all grew up in Houston and have lived here for all of their lives. To help you keep track of who you are listening to, you will see a picture of each person. We asked each of them to read a list of words, for example "cap," "window," etc. You will hear words from these recordings.

After the 16 experimental blocks the experiment ended with two open response questions to which participants responded by typing their answers into a text box. The first question asked how difficult the participants found the task. It was explained that this might include, for example, the speed at which the words were presented or whether some words were more difficult than others.

This question was designed to ensure that the task was relatively easy without becoming exceedingly monotonous in the light of the large number of trials. Pilot versions of the experiment had been rated as too slow. The second question asked whether there was anything about the different speakers' voices or accents that struck the participants as interesting or unusual. This question was primarily designed to test the success of the speech synthesis given that all experimental vowels were synthesized (see below). A second objective was to evaluate the success of creating congruous an incongruous speaker-dialect pairings in so far as this is reflected in overt commentary.

The experiment was followed by a verbal debriefing in which the participants were given the opportunity to ask questions and learn about the research hypotheses. The debriefing was intended to be unconstrained and conversational, so as to bring out any unanticipated reactions which would later help in interpreting the results. Before revealing the fact that the identities of the speakers were fictitious and that the different pronunciations were synthetically produced, I tried to elicit the participants' general impression of the speakers they had heard. A good conversation starter turned out to be the question

whether the pronunciations the participants had heard were familiar to them as native Houstonians. If yes, I asked what those pronunciations were. If no, I asked how the pronunciations they had heard were different from those which they considered more typical of Houston. I also tried to obtain each participant's impression of any differences between the two members of the perceived age comparison and the perceived ethnicity comparison.

The experiment was administered entirely by the author, who is a near-native speaker of non-Southern Anglo American English. Participants at the University of Houston-Downtown performed the task in a quiet conference room. Participants at Rice University performed the task in a soundproof booth. The experiment took about 25 minutes to complete.

3.3 Lexical items

Each of the four vowels under investigation was represented by two monosyllabic words. The stimulus words were common nouns and adjectives.

They were chosen to form two /eɪ/-/ɛ/ pairs and two /ʌ/-/ɑ/ pairs each sharing the same onset consonant. The pairs are shown in Table 3.3.

/eɪ/-/ɛ/ pairs		/ɑ/-/ʌ/ pairs	
/eɪ/	/ɛ/	/ɑ/	/ʌ/
bay	*bed*	*dock*	*duck*
day	*dead*	*stock*	*stuck*

Table 3.3: Pairs of lexical items heard as experimental stimuli

3.4 Visual stimuli

The fictitious names and ages, as well as the fictitious regional labels displayed together with the photo at the beginning of each block are shown in Table 3.4.

Practice speaker	Older Anglo male	Younger Anglo male	Young AfrAm female	Young Anglo female
Daniel, 19	Clint, 58 from SE Houston	Michael, 33 from NW Houston	Alicia, 18 from Missouri City	Emily, 18 from Katy

Table 3.4: Fictitious names, ages, and regional labels

The purpose of providing explicit ages in addition to the age information that could be inferred from the photographs was to control the variable of perceived speaker age more precisely. This was done at the risk of drawing attention to the fact that speaker age was being studied. The ages of 33 and 58 assigned to the male speakers were a compromise between the production data discussed in Chapter 2 and the constraints of the matched-guise design. Judging by the Anglo production data, speakers in their early 30s and speakers in their late 50s may show a substantial dialect difference in Houston, so that these two ages create a large enough age contrast. At the same time, a 25-year age difference was deemed small enough in order for a single voice to pass sufficiently well as a speaker of either age group.

Similarly, for the two female speakers the age of 18 was a compromise between the production results and the practical requirements of the speech synthesis process. Judging by the production data discussed in Chapter 2, the youngest teenage African-Americans show the greatest dialect difference relative to their same-age Anglo peers. Therefore, the incongruity should be the greater

the lower the perceived teenage speakers' age is. The age of 18 was deemed sufficiently young yet old enough to assume that the speaker has an adult vocal tract. This, in turn, allowed the design of the auditory stimuli to be based on reference studies of adults, and to avoid complications having to do with adolescent voice characteristics.

The purpose of providing a proper name and a regional label for each speaker was to reinforce in the minds of the participants the idea that the speakers were real Houstonians, as well as to reinforce the concept of Houston in general. The names and regional labels were chosen from the real first names and real places of residence of recorded HUES speakers in the relevant age and ethnicity categories in the HUES sample.

3.5 Auditory stimuli

The two members of each auditory stimulus pair, e.g. the words *bay* and *bed*, were each heard 12 times per block. These 12 trials were further subdivided such that three variants of each vowel were heard. For example, participants

heard the word *bay* spoken with three dialectal variants of /eɪ/, ranging from most to least Southern. Each vowel variant was heard four times. I will refer to the three variants of each vowel as, for example, /eɪ/-1, /eɪ/-2, and /eɪ/-3. The point of interest in the experiment was the participants' response to the more "extreme" variants, variants 1 and 3, in interaction with one of the different speaker guises, as explained above. The intermediate variants served as fillers. Their function was to distract the participants' overt attention from the properties of variants 1 and 3. For the vowels /eɪ/, /ɛ/, and /ʌ/, the numbering follows the increase in the first formant, so that the most raised (or, least lowered) variants receive the number 1 and the most lowered (or, least raised) variants receive the number 3. For /ɑ/, the numbering follows the increase in the second formant, so that the most backed (or, least fronted) variant receives the number 1, and the most fronted (or, least backed) variant, receives the number 3.

The four repetitions of each vowel variant were further differentiated such that each vowel was synthesized with a different *f0* contour. This was done to make the stimuli more realistic by introducing natural variability. The four

contours differed in that two of them ended in a rise, while the other two did not. All four *f0* contours used in the experiment showed some fluctuation because they were extracted from naturally produced words (see below).

3.6 Auditory stimulus creation

The goals of the auditory stimulus creation process were dictated by two requirements. First, in terms of dialectal variation, the stimulus words had to exhibit the spectral features described in Chapter 2, viz. Southern and non-Southern Anglo variants of /eɪ/ and /ɛ/, and Anglo and African-American variants of /ɑ/ and /ʌ/. Second, the matched-guise design required the stimuli to sound not only natural but appropriate in terms of voice quality for speakers of the relevant social categories.

3.6.1 Formant trajectories

In order to create realistic and representative formant shapes, the stimuli were based as closely as possible on recorded vowels spoken by HUES speakers with the relevant dialect features. In the case of the /eɪ/ and /ɛ/ words, this was relatively easy to achieve because three of the four lexical items, *bay, day*, and *dead*, were contained in the HUES word list. The formant shape of the fourth item, *bed*, was taken from the HUES word *Fed*, with minor adjustments to the slope of the in-transitions of the first two formants to improve naturalness. In the case of the words *duck, dock, stuck*, and *stock*, none of which were contained in the HUES word list, the formants were manually constructed by closely copying and combining the beginning and end of the formants of HUES words which are identical or similar in place of articulation to the stimulus words: *box, puck*, and *not*.

The procedure for constructing variants 1 and 3 of each vowel was as follows. First, for each word the five HUES speakers who show the greatest amount of raising, lowering, fronting, or backing of the relevant vowel in the

relevant phonological environment were identified by inspecting F1-by-F2 plots of the position of the vowel nucleus of the relevant words. The F1, F2, and F3 trajectories of the five word tokens were then extracted, using the procedure of extracting formant contours described in Chapter 2. The extracted formant frequency values were then normalized using the each speakers' Nearey-1 scaling factor. Next, average normalized F1, F2, and F3 trajectories were calculated over these five speakers by calculating the mean formant frequency at each of the 100 measurement points. Finally, normalized and averaged trajectories were linearly scaled up or down, so that their overall position in F1-by-F2 vowel space was in a place appropriate for an adult male or female vocal tract. This was done by first mapping the average vowel space for adult speakers of American English in three reference studies: Peterson & Barney (1952), Hillenbrand, Getty, Clark & Wheeler (1995), and Hagiwara (1997). The Lobanov-normalized vowel means of the HUES speakers were superimposed on and aligned with this average vowel space. This allowed the formant frequencies of the stimuli to be scaled to a position which precisely reflects the HUES results as well as being typical of an adult male or adult female vocal tract.

Variant 2 of each vowel was designed to have formant frequency values that are perceptually intermediate between variants 1 and 3. To produce these variants, at each point along the trajectory of variants 1 and 3 the F1, F2, and F3 values were converted to Bark, using the *hertzToBark* function in Praat, averaged, and converted back to Hertz using Praat's *barkToHertz* function. The resulting formant contours of variant 1 and variant 3 are shown in Figures 3.2 to 3.7. In each case, the left panel shows the F1, F2, and F3 values across normalized time and the right panel shows the trajectory of the first two formants in F1-by-F2 vowel space.

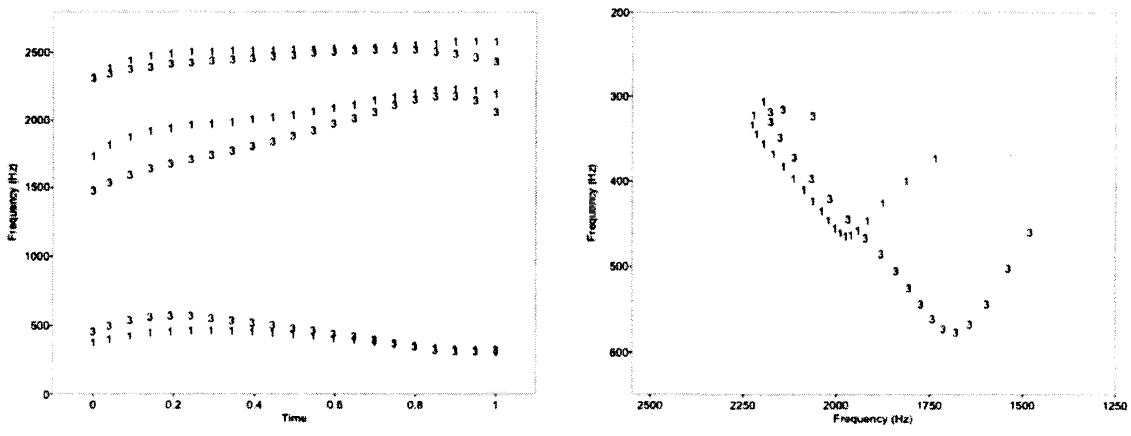

Figure 3.2: Formant contours of /eɪ/ in *bay*

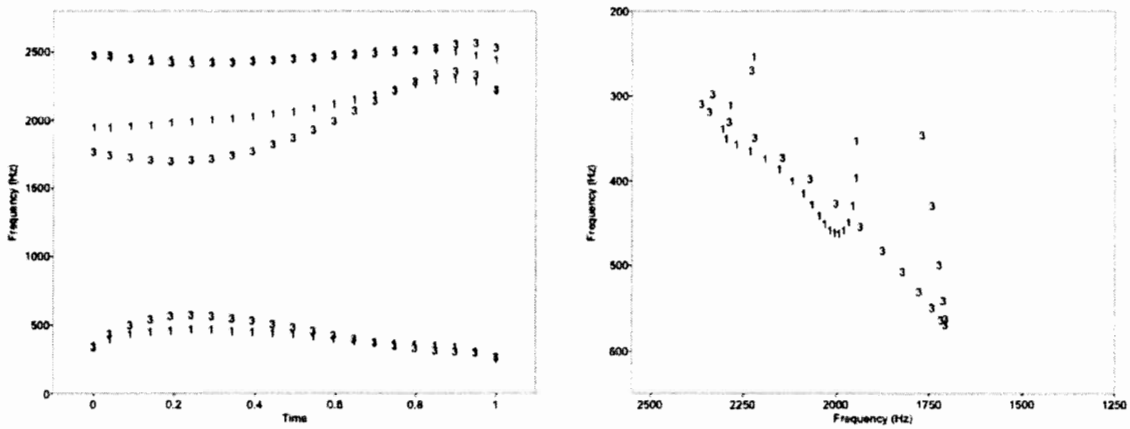

Figure 3.3: Formant contours of /eɪ/ in *day*

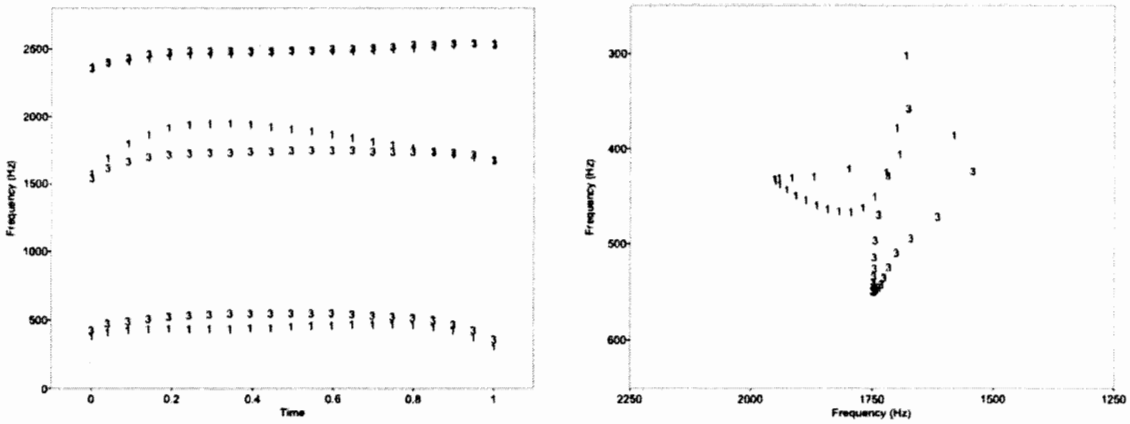

Figure 3.4: Formant contours of /ɛ/ in *bed*

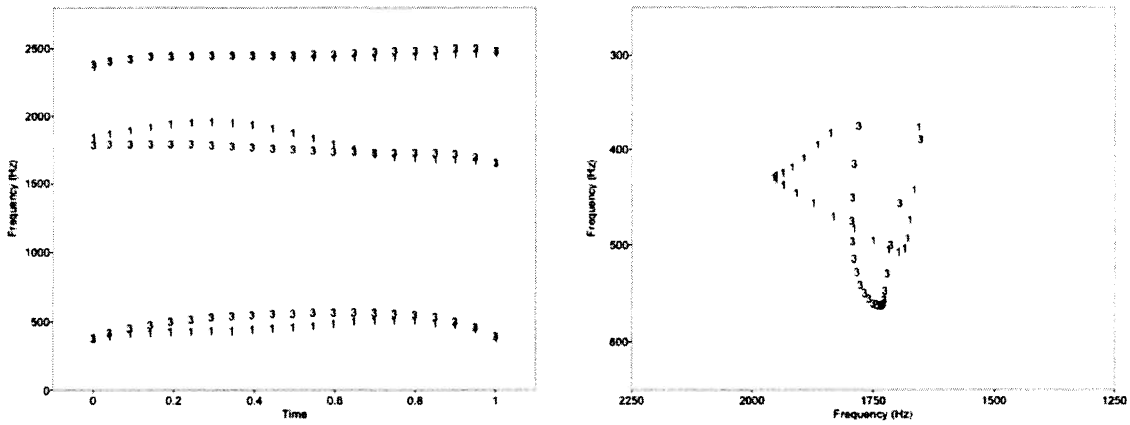

Figure 3.5: Formant contours of /ɛ/ in *dead*

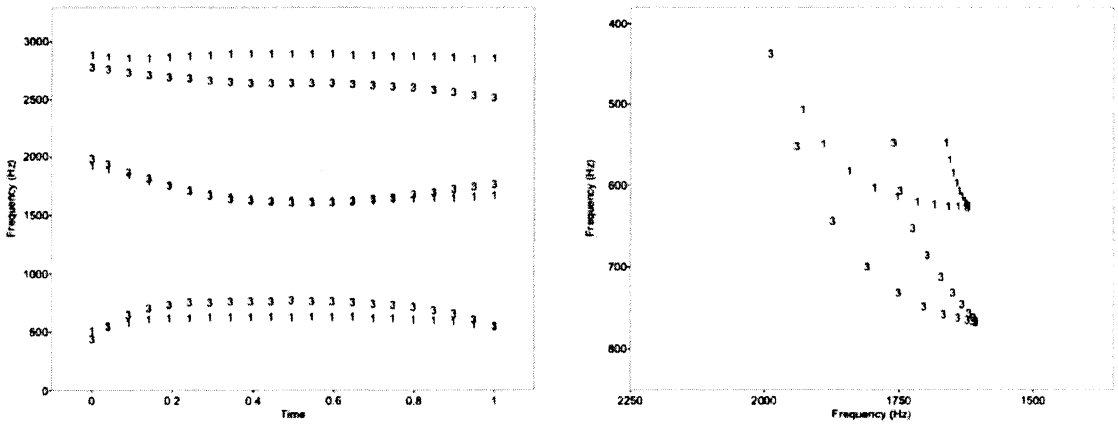

Figure 3.6: Formant contours of /ʌ/ in *duck* and *stuck*

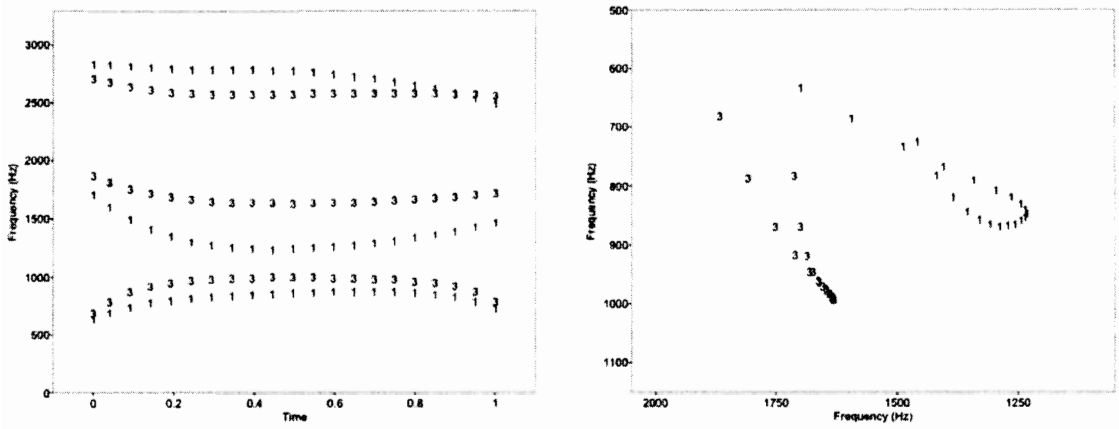

Figure 3.7: Formant contours of /ɑ/ in *dock* and *stock*

3.6.2 Pitch contours

Two sets of four *f0* contours were used to synthesize the /eɪ/ and /ɛ/ words as well as the /ɑ/ and /ʌ/ words. To ensure naturalness, the *f0* contours were extracted from recordings of HUES speakers of the relevant gender using Praat's default pitch tracking settings. The contours used for /eɪ/ and /ɛ/ were chosen from tokens of *day* and *dead* spoken by middle-aged male Anglo speakers. The contours of /ɑ/ and /ʌ/ were chosen from readings of the word *box* by female teenage speakers, both Anglos and African-Americans. The original *f0* values were linearly shifted upwards or downward so that the pitch mean provided by

the function *Get mean (curve)...* was 115 Hertz for the male speakers and 195 Hertz for the female speakers. The contours in their final form are shown in Figure 3.8 below.

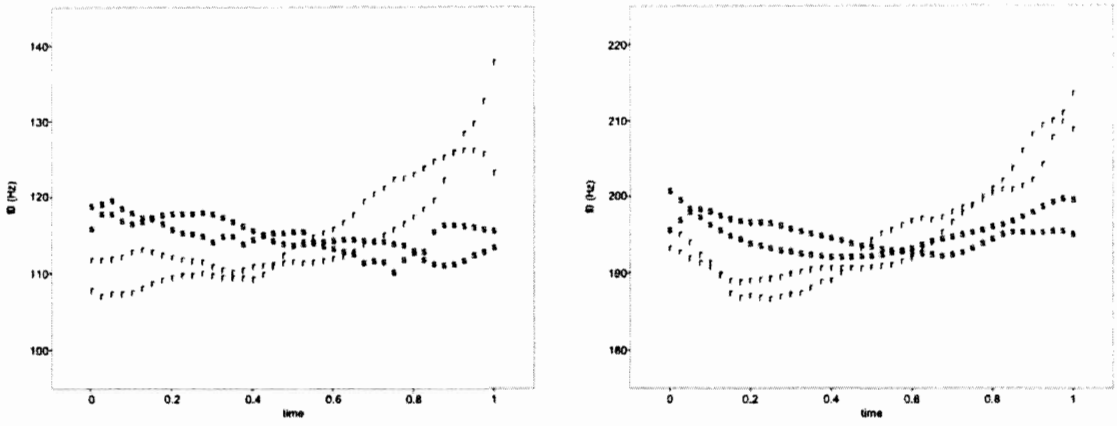

Figure 3.8: Pitch contours of /eɪ/ and /ɛ/ (left), and /ʌ/ and /ɑ/ (right). "r" = rising, "s" = non-rising

3.6.3 Vowel duration

For /ʌ/ and /ɑ/, the duration of the vowel was held constant across the three variants. In the case of both /eɪ/ and /ɛ/, it was varied slightly. The duration values are shown in Table 3.5.

Vowel	Variant 1	Variant 2	Variant 3
/eɪ/	300 ms	320 ms	340 ms
/ɛ/	230 ms	245 ms	260 ms
/ʌ/	159 ms	159 ms	159 ms
/ɑ/	170 ms	170 ms	170 ms

Table 3.5: Durations of the three variants of each vowel

The values in Table 3.5 are based on the durations of the relevant vowels in the HUES wordlist recordings. Duration measurements showed no significant difference between the young Anglo and African-American speakers for /ʌ/ and /ɑ/ in the relevant phonological context. Therefore, duration was held constant for these vowels. For Anglo /eɪ/ and /ɛ/, however, vowel duration was increased with the degree to which the vowel was Southern-shifted, as would be expected in Southern American English (Feagin 1987, Thomas 2003, Wetzell 2000). Therefore, the duration was set to increase with the degree of raising of /ɛ/ and the degree of lowering of /eɪ/. The exact increments were not strictly based on the production data, but rather reflect a compromise between the duration values in the production data and considerations of naturalness.

3.6.4 Synthesis procedure

The vocalic parts of each stimulus word were produced using the KlattGrid formant synthesizer (Weenink 2009) implemented in recent versions of Praat. Formant synthesis was chosen despite the drawbacks that this form of synthetic speech has historically been shown to have (Pisoni, Nusbaum, Luce & Slowiaczek 1985, Duffy & Pisoni 1992; but see Pisoni 1997 for a more optimistic outlook). The main reason for using synthesis is the degree of control over the acoustic parameters that it allows. Only the vowel portions of each stimulus word were synthesized. Onset and coda consonants were taken from actually recorded words in the HUES word list recordings which were spliced onto the synthetic vowels.

The synthesis procedure was a two-step process. In the first step, a naturally produced, recorded vowel was copy synthesized in as much spectral detail as possible. The aim of this step was to arrive at a synthetic vowel with a known, natural voice quality. The speakers whose voices were copied were two

male Anglos, one in his 40s and the other in his 50s, a female teenage African-American speaker and a female teenage Anglo speaker, all recorded for the HUES project. In the second step, the first three formants of the copy were replaced by the formant trajectories described in Section 3.6.1, the pitch contour was replaced with one of the pitch contours described in Section 3.6.2, and the duration of the vowel was adjusted as described in Section 3.6.3.

The first step worked as follows. A Praat script was written which extracts the recorded vowel's fundamental frequency contour, formant frequencies and bandwidths, as well as the original amplitude contour, and combines them into a single KlattGrid object. This script resembles Praat's *To KlattGrid...* function. One difference between it and *To KlattGrid...* is that the latter does not allow LPC values that are not even integers. Next, the formants were manually edited to remove mistracked points. Such points, along with the associated bandwidth values, were removed and automatically interpolated. The first four or five formants were extracted from the original, depending on whether the fifth formant was trackable. Higher formants were automatically generated and given default formant and bandwidth values. The intensity contour was manually

edited where necessary to erase unnatural effects most likely due to the way in which source and filter are combined in the synthesis process. Typically, the contour was smoothed.

The resulting KlattGrid object was then manually fine-tuned to find the best possible auditory match to the original speaker's voice quality. This was done following Alwan, Bangayan, Gerratt, Kreiman & Long's (1995) spectral matching procedure. An interactive Praat interface was written for this purpose, based on Praat's demo window function. This interface allowed all KlattGrid parameters, especially the voice quality parameters, to be adjusted at 20 equidistant steps through the vowel while providing instant auditory feedback so that any changes can be evaluated and, if necessary, reversed. The following adjustments were typically made to the synthetic vowels: (i) boosting the amplitude of higher formants, especially in the case of the male speakers; (ii) manually increasing or decreasing bandwidth values in order to eliminate artificial amplitude dips and spikes; (iii) adding breathiness noise, both to female and male voices; (iv) increasing spectral tilt at the edge of vowels

preceding voiceless stops to model the appropriate consonant-vowel co-articulation (Ní Chasaide & Gobl 1993).

3.7 Participants

The choice of participants in the experiment was crucial in the light of the experience effects discussed in Chapter 1. The goal was to recruit participants who possess the sociophonetic knowledge at issue in the experiment. In practical terms, this was defined as listeners who had grown up and lived for most or all of their lives in the Houston metropolitan area. 60 listeners participated in the experiment, 23 at the University of Houston-Downtown (UH-D) and 37 at Rice University. The UH-D participants were recruited in two ongoing English classes and received extra course credit for their participation. The participants at Rice University were recruited using on-campus announcements seeking native Houstonians for a speech perception experiment. They were paid $10 for their participation. Most of the Rice participants were undergraduate and graduate

students, but some were university employees or Houston community members who saw the experiment advertised on campus.

Data from two of the 37 Rice participants were excluded from the analysis. One participant had lived abroad for most of his adult life. The other excluded participant had answered 'yes' to the question asking whether she was currently experiencing speech or hearing problems. Data from 10 of the 23 UH-D participants also had to be excluded. Five participants were not native speakers of English but of Spanish, according to the definition used here (see Section 3.2). Four participants had not grown up in the Houston metropolitan area. Data from one participant was excluded because equipment failure resulted in the loss of data points. In total, data from 48 participants was analyzed quantitatively.

All 48 participants whose data was included in the quantitative analysis were either born in Houston or had moved to Houston before turning 7 years of age. All had lived in Houston continuously between age 6 and 18, except for one participant who had lived in France for 1 year, one who had lived in Columbus, Ohio for 3 years, and one who had lived in Tulsa, Oklahoma for 2 years. 37

participants had lived only in Houston after age 18, while the other 11 participants had lived away from Houston for between 1 and 5 years. 14 participants reported speaking another language besides English natively (7 Spanish, 2 Mandarin, 2 Vietnamese, 1 Korean, 1 Hindi, 1 Yoruba). None reported experiencing speech or hearing problems.

Participants were randomly assigned to either Group A or Group B. The age, gender, and ethnicity distribution of each group is shown in Table 3.6.

	Group A	Group B
Total number	n = 24	n = 24
Age in years	range 19–59, mean 27.9, median 22.5	range 18–45, mean 24.9, median 22
Gender	19 female, 5 male	20 female, 4 male
Ethnicity	8 African-American, 6 Anglo, 6 Asian, 4 Hispanic	8 Anglo, 6 African-American, 5 Hispanic, 5 Asian

Table 3.6: Participant demographics

The participants' age was calculated by subtracting the self-reported year of birth from 2010. Participant ethnicity was determined by categorizing the self-reported ethnicity/race label into four larger groups, as seen in Table 3.6. The

'African-American' group includes the actual labels *Black, Black American, African American* and *Nigerian.* The 'Anglo' group includes *White* and *Caucasian.* The 'Hispanic' group includes *Hispanic, Mexican American,* and *Mexican.* The 'Asian' group includes *Asian, Asian-American, Vietnamese-American, Spanish-Filipino* and *South Asian.* In cases where participants self-identified as belonging to two of these categories (e.g., "half Hispanic, half White", "Black/White"), the category assignment was based on the first of the two designations given.

Chapter 4

4. Results

In this chapter I present the quantitative and qualitative results of the speech perception experiment described in Chapter 3. I begin with a summary of the participant's response accuracy (Section 4.1). The core of the quantitative analysis is formed by two linear mixed effects regression models which were fit to the response time data gathered in the two parts of the experiment: the comparison of the participants' responses under the perceived age manipulation and the comparison of the participants' responses under the perceived ethnicity manipulation. I will refer to these two parts as the *perceived age comparison* and the *perceived ethnicity comparison*, respectively. At other times, I will refer to the two parts with reference to the vowels whose quality were manipulated. Thus, in the former case I will refer to the */eɪ/ and /ɛ/ trials*, and in the latter case I will

refer to the */a/ and /ʌ/ trials*. The two regression analyses are presented in Section 4.2. The participants' feedback is summarized in Section 4.3. The results presented in this chapter will be further discussed in Chapter 5 and Chapter 6.

4.1 Response accuracy

The proportion of correct responses out of all 18432 recorded responses was 97.0%. Incorrect responses occurred in 2.8% of the trials. In addition, there were 0.2% non-responses, i.e., trials in which no response occurred within 3 seconds of the onset of the sound file. Given that the task was not designed to create difficult processing conditions, but in fact to avoid phonological ambiguity, the high rate of accurate responses is not surprising. The small number of errors suggests that the experiment was successful in creating a task that is easy to perform. This is also reflected in the participants' feedback comments, which are discussed below.

A closer look at the distribution of the incorrect responses suggests that the actual error rate was probably even lower. Manual inspection revealed what

appear to be at least two different error types. Occasionally, the pattern in which incorrect responses occurred suggests that the participants were in fact making accurate decisions but failed to remember the mapping of the response alternatives to the left and right button. Recall that the mapping was given to the participants only at the start of each 24-trial block. They had to keep the mapping in memory for the duration of each block. Also, the mapping was reversed in the second half of the experiment. It appears that as a result of this potential source of difficulty, participants sometimes consistently responded by pressing the wrong button. This is most clearly seen in extended, uninterrupted sequences of incorrect responses, especially those starting at some point within a block and continuing until the end of that block. As discussed below in Section 4.3, some of the participants' feedback comments suggest that this is indeed what happened. This type of incorrect response is, then, qualitatively different from other types of error and has to be separated out.

In order to estimate the amount of these systematic errors, all incorrect responses which occurred in direct sequence with another incorrect response or a non-response were discarded. This leads to a reduction of the rate of incorrect

responses to only 1.7% overall and an increase of the accurate response rate to 98.1%. Given this very low number of inaccurate responses, and given that no specific hypothesis about response accuracy was formulated in Chapter 2, no further analysis of the response accuracy data was carried out.

4.2 Response time

The predicted effect of sociophonetic congruency was an effect on to the speed of accurate word recognition. Therefore, the analysis of response time (henceforth, RT) was based only on data from trials in which participants responded correctly. Prior to the analysis, all responses to filler trials were discarded, which reduced the number of data points by about one third. Also discarded were responses with recorded RTs of less than 500 ms. This was based on the following consideration. The auditory stimuli were designed so that the vowel onset occurred after exactly 200 ms. In addition, the muscular response associated with pressing a button was estimated to incur a 200-300 ms delay. Taken together, this means that at a point 500 ms after the onset of the sound

file at least a minimal portion of the vowel can be assumed to have been processed so that responses which must have been guesses are excluded. Excluding RTs lower than 500 ms resulted in a reduction of the remaining data by 3.1%. Further excluded were recorded RTs much larger than the median RT so as to prevent skewing of the overall distribution by outliers. This was done separately for the data from the perceived age comparison (i.e., the /eɪ/-/ɛ/ trials) and the data from the perceived ethnicity comparison (i.e., the /ʌ/-/ɑ/ trials). For each set, a cut-off point of twice the standard deviation above the median was calculated, and all values exceeding this point were discarded. This resulted in a reduction of the remaining /eɪ/-/ɛ/ data by 4.0% and a reduction of the remaining /ɑ/-/ʌ/ data by 4.5%.

In order to test whether the predicted interaction effect between vowel variant and speaker guise is borne out by the RT data, all remaining trials were coded as either 'congruent' or 'incongruent' depending on the pairing of vowel variant and speaker guise. In other words, the interaction between the variable vowel variant and the variable speaker guise was coded as a simple, binary

variable. This was done in order to make it easier to test further interactions of the variable of congruency with other variables.

The data from the perceived age comparison and the data from the perceived ethnicity comparison were analyzed separately. In each case, a linear mixed-effects regression model was hand-fit to the RT data following the same model fitting and significance testing procedure as in the analysis of the formant frequency data in Chapter 2. Two predictors were included as random effects: the identity of the participant and the particular word that the vowel occurred in. All other variables tested were entered as fixed effects. In the following discussion, all such variables are shown in small capitals, for example the variable CONGRUENCY.

Whenever CONGRUENCY made a significant contribution to the regression model, either by itself or in interaction with another variable, the variables VOWEL VARIANT (e.g., Southern or non-Southern variant) and SPEAKER GUISE (e.g., older or younger speaker photo) were also included in the model, even if they did not make a significant contribution by themselves. This is because CONGRUENCY is defined as the interaction of the two.

In addition to testing for a main effect of CONGRUENCY, possible main effects of VOWEL VARIANT and SPEAKER GUISE and possible interactions between CONGRUENCY, VOWEL VARIANT or SPEAKER GUISE and a range of other variables were also tested. These other variables can be divided into three sets. The first set includes variables pertaining to the position of a trial in the experiment: the BLOCK in which the trial occurred (from 1 to 16), the position of the trial in the current block, or TRIAL IN BLOCK (from 1 to 24), and whether the trial occurred in the first or in the second half of the experiment, or EXPERIMENT HALF. The second set are variables pertaining to properties of the auditory stimuli besides the variable VOWEL VARIANT. They are the VOWEL that was heard (e.g., /eɪ/ or /ɛ/) and the stimulus VOICE (e.g., male-1 or male-2). The third set are variables pertaining to the participants. They are PARTICIPANT AGE, PARTICIPANT GENDER, and PARTICIPANT ETHNICITY. In order to arrive at a maximally complete model, main effects of each of these additional variables and pair-wise interactions between them were also tested.

In the discussion of the regression models below, the variable PARTICIPANT ETHNICITY is not initially included. Instead, effects of PARTICIPANT ETHNICITY on

each data set were tested in a separate, subsequent step. The reason for proceeding in this way is that, unlike all other variables, PARTICIPANT ETHNICITY takes on four non-continuous values in the categorization scheme used here: 'Anglo', 'African-American', 'Asian', and 'Hispanic'. Each ethnicity comparison was based on a four-fold re-coding of this variable so as to make it binary. For example, a possible effect of Anglo ethnicity was tested by adding to the fitted model the variable of ethnicity coded as 'Anglo' and 'non-Anglo'. Thus, I will speak of four different variables: ANGLO, AFRICAN-AMERICAN, ASIAN and HISPANIC. Because this analysis of ethnicity is a four-fold comparison, the significance level for all ethnicity effects was lowered to 0.0125 (0.05 divided by four). The results of the four ethnicity analyses will be discussed separately below, following the summary of all other fixed effects.

To verify the robustness of the congruency effects in the full models presented below, simpler models were also constructed. These models were built using the same random predictors but as fixed effects only the variables CONGRUENCY, VOWEL VARIANT and SPEAKER GUISE as well as the interaction effects

which CONGRUENCY enters into. In all cases, the congruency effects remained significant.

4.2.1 Perceived age comparison

The fixed effects output of the regression model of the perceived age data is shown in Table 4.1.

Predictors of RT	Estimate	Std. error	*t*-value
Congruency Incongruent	12.4943	6.0396	2.069*
Variant Southern	90.7217	11.9769	7.575***
Speaker guise Younger	-6.2628	3.1918	-1.962*
Block	-5.0241	0.7049	-7.127***
Trial in block	1.3369	0.3712	3.602***
Experiment half Second	21.8137	5.8129	3.753***
Vowel /eɪ/	3.6264	13.9416	0.26
Participant age	2.45	0.9618	2.547*
Participant gender Male	-2.8178	22.176	-0.127
Congruency Incongruent : Trial in block	-1.1194	0.4217	-2.654**
Variant Southern : Block	1.2232	0.6196	1.974*
Variant Southern : Trial in block	-1.9789	0.4266	-4.639***
Variant Southern : Vowel /eɪ/	-45.2593	5.691	-7.953***
Variant Southern : Age	-0.8602	0.3092	-2.782**
Speaker guise Younger : Part. Gender Male	18.2698	7.2188	2.531*
Participant age : Vowel /eɪ/	1.7453	0.3092	5.645***

Table 4.1: Fixed effects in the regression model fit to the perceived age data. Symbols following the *t*-value indicate the associated *p*-value: '***' $p<0.001$, '**' $p<0.01$, '*' $p<0.05$, '.' $p<0.1$

I begin the discussion of the large number of significant effects in Table 4.1 by discussing and illustrating the effects involving sociophonetic congruency. As predicted, there is a main effect of CONGRUENCY, in the predicted direction. By itself, this effect suggests that congruous trials were indeed recognized more quickly than incongruous trials ($p=0.039$). However, this main effect has to be

interpreted in the context of the additional interaction effect of CONGRUENCY and TRIAL IN BLOCK ($p = 0.008$). The direction of the interaction effect is such that the inhibitive effect of CONGRUENCY changes over time, becoming weaker in the course of the 24 trials in each block. The more trials participants had completed within a block, the more quickly they responded to incongruous trials or the more slowly they responded to the congruous ones. In fact, adding up the regression coefficients of the main effect and the interaction effect in Table 4.1 suggests that at the end of each block the main effect of CONGRUENCY was more than offset by the interaction effect so that at this point the participants' responses to incongruous trials were *faster* than their responses to congruous trials. This crossover can be seen in the raw data. The mean RT values across the 24 trials per block for both congruous and incongruous /eɪ/-/ɛ/ trials are illustrated in Figure 4.1. The data points plotted in Figure 4.1 are the raw RT means and standard errors, except that each raw RT value was adjusted by adding or subtracting from it the predictions of the random effects of PARTICIPANT and WORD yielded by the mixed model. In order to visualize the interaction

effect, linear trend lines were added by fitting linear regression lines to the congruent and incongruent RT sets.

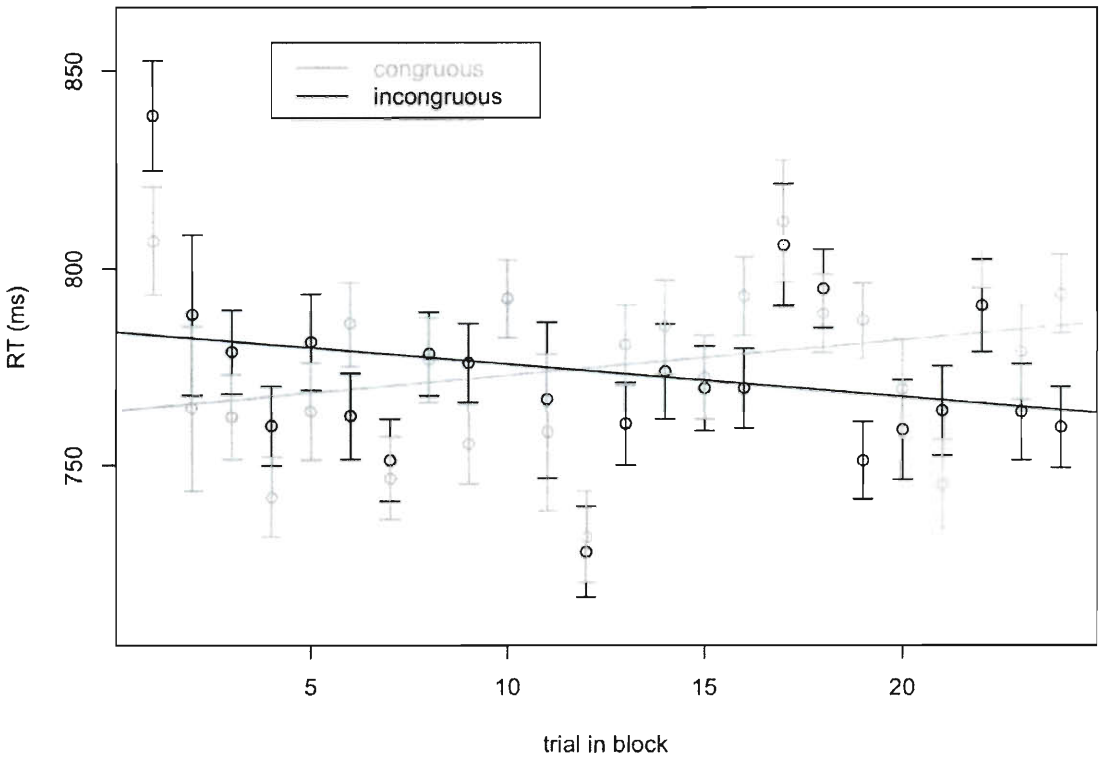

Figure 4.1: Mean RT in /eɪ/ and /ɛ/ trials across the 24 trials per block

The interaction of CONGRUENCY and TRIAL IN BLOCK seen in Figure 4.1 is such that there is an initial advantage of congruous trials. The first trial per block is responded to most slowly overall, but more importantly, there is a clear RT

difference between congruous and incongruous trials. Comparing the RT difference between incongruous and congruous trials across all 24 positions shows that in the course of the block this advantage is indeed reduced and actually reversed near the midpoint of the block. At the end, it is the congruous trials which have an RT advantage.

A final point regarding the variable CONGRUENCY is that it does not interact with either of the other two variables pertaining to the position of a trial in the experiment, BLOCK and EXPERIMENT HALF. This suggests that the effect of CONGRUENCY was variable only within each block but did not change overall across the experiment.

Coming to the other significant effects in the regression model in Table 4.1, there are main effects of all three variables pertaining to the temporal position of a trial within the experiment: TRIAL IN BLOCK, BLOCK, and EXPERIMENT HALF. By itself, the first effect suggests that response times slowed down slightly within a block. The second effect suggests that, at the same time, participants responded more quickly as they completed more blocks. The third effect suggests that responses in the second half of the experiment were generally

slower than responses in the first half. In each case, the main effect has to be interpreted within the context of the additional interaction effects of these three variables.

Next, there is a strong main effect of VOWEL VARIANT. Note that, judging by its regression coefficient, this effect is by far the largest predictor of RT in the model. Words containing the Southern vowel variants, i.e. a raised /ɛ/ or a lowered /eɪ/, were recognized considerably less quickly overall. This variable also shows a number of interactions. The interaction with VOWEL CATEGORY shows that the effect of VOWEL VARIANT was stronger for /ɛ/ than for /eɪ/. The raised /ɛ/ had, judging by the regression coefficients, twice the inhibitive effect of the lowered /eɪ/. The position of the trial in a block and the listener's age both reduced the inhibitive effect of the Southern vowel variants as seen in the interaction of VOWEL VARIANT with both TRIAL IN BLOCK and PARTICIPANT AGE. The more trials per block had been completed, the faster listeners responded to the Southern variants. Older participants were faster to recognize Southern vowel variants than younger participants. Note, however, that older participants were overall slower to respond to all trials as seen in the main effect of AGE. Finally,

there is an interaction of VOWEL VARIANT and the variable BLOCK. Taken together with the main effect of BLOCK, this interaction suggests that the increase in the speed with which participants responded as they completed more blocks was smaller for the Southern variants. However, overall there was an accelerating effect across blocks, even for the Southern variants.

The last main effect in the model is of the variable SPEAKER GUISE. Trials in which participants saw the photo of the younger male Anglo were responded to faster overall. However, the interaction of SPEAKER GUISE and PARTICIPANT GENDER suggests that male participants responded more slowly than female listeners to the younger speaker and more quickly to the older speaker. Judging by the regression coefficients of these two effects, the interaction effect outweighs the general advantage of seeing the younger male Anglo.

The lack of a main effect of VOWEL CATEGORY shows that whether participants heard a word containing /eɪ/ or a word containing /ɛ/ did not, by itself, affect their response times. Instead, there is an interaction of VOWEL CATEGORY and PARTICIPANT AGE such that younger participants had an advantage relative to older participants in recognizing /eɪ/ and a disadvantage in

recognizing /ɛ/. Note that this is independent of which variant of /eɪ/ or /ɛ/ was heard, and also in addition to the advantage of older speakers in recognizing the Southern variants of these vowels, as noted above.

Adding the variable PARTICIPANT ETHNICITY to the complete model in the stepwise fashion described at the beginning of this section never yielded a main effect of PARTICIPANT ETHNICITY or an interaction effect of PARTICIPANT ETHNICITY with CONGRUENCY. Thus, there is no evidence that any ethnically defined subgroup of participants responded more quickly or more slowly overall than all others, and no evidence that any such group responded more quickly or more slowly than all other groups depending on whether the trial was congruent or incongruent. Nevertheless, for three of the four participant ethnicity groups there were additional interaction effects or changes in the significance of existing predictors.

First, adding the variable ANGLO to the model in Table 4.1 yields three additional interaction effects. These effects and the non-significant main effect of ANGLO are shown in Table 4.2.

Additional predictors of RT	Estimate	Std. error	*t*-value
Ethnicity Non-Anglo	-23.6721	21.3966	-1.106
Ethnicity Non-Anglo : Trial in block	1.1657	0.4575	2.548*
Ethnicity Non-Anglo : Block	2.5879	0.686	3.772***
Ethnicity Non-Anglo : Vowel /eɪ/	16.3937	6.3105	2.598**

Table 4.2: Changes in the regression model fit to the perceived age data when the variable ANGLO is added

The first interaction is between ANGLO and TRIAL IN BLOCK. It goes along with a change in main effect of TRIAL IN BLOCK. When the interaction between ANGLO and TRIAL IN BLOCK is added, TRIAL IN BLOCK no longer shows a significant main effect ($p = 0.32$) This suggests that the participants' slowing down in the course of each block is carried entirely by the non-Anglo participants. The second interaction is between ANGLO and BLOCK. This interaction effect goes along with a change in the coefficient of the main effect of BLOCK from -5.02 to -6.88. Together, these two changes suggest that while all participants came to respond faster as they completed more blocks, the accelerating effect was greater for the Anglo participants than for the non-Anglo participants. The third interaction is between ANGLO and VOWEL. Anglo participants responded more quickly than non-

Anglo participants to the vowel /eɪ/. As previously, there is no main effect of VOWEL.

Adding the variable AFRICAN-AMERICAN to the model in Table 4.1 yields one additional interaction effect but no changes in the significance of existing predictors. The additional interaction and the non-significant main effect of AFRICAN-AMERICAN are shown in Table 4.3.

Additional predictors of RT	Estimate	Std. error	*t*-value
Ethnicity Non-African American	27.0245	20.9047	1.293
Ethnicity Non-African American : Block	-1.9600	0.6802	-2.882**

Table 4.3: Changes in the regression model fit to the perceived age data when the variable AFRICAN-AMERICAN is added

The interaction effect goes along with a change in the coefficient of the main effect of BLOCK from -5.02 to -3.63. Together, these two changes suggest that unlike the other participants the African-American participants did not come to respond faster as they completed more blocks.

Finally, adding the variable HISPANIC to the model in Table 4.1 yields one additional interaction effect but no significant changes in the role of the other

predictors. The interaction effect and the non-significant main effect of HISPANIC are shown in Table 4.4.

Additional predictors of RT	Estimate	Std. error	*t*-value
Ethnicity Non-Hispanic	-20.9971	21.7781	-0.964
Ethnicity Non-Hispanic : Variant Southern	-26.3035	7.2544	-3.626***

Table 4.4: Changes in the regression model for the perceived age comparison when the variable HISPANIC is added

The interaction between HISPANIC and VARIANT suggests that the Hispanic participants were slower to respond to the Southern vowel variants than the non-Hispanic participants.

4.2.2 Perceived ethnicity comparison

The fixed effects output of the regression model fit to the perceived ethnicity comparison is shown in Table 4.5.

Predictors of RT	Estimate	Std. error	*t*-value
Variant Anglo	30.0911	2.8732	10.473***
Speaker guise Anglo	5.9743	2.8683	2.083*
Trial in block	0.5452	0.2997	1.819.
Vowel /ʌ/	-13.7486	20.444	-0.673
Block	-3.5006	0.6413	-5.459***
Experiment half Second	32.2987	5.8561	5.515***
Participant age	1.9453	0.8612	2.259*
Vowel /ʌ/ : Trial in block	-1.8702	0.4258	-4.392***
Vowel /ʌ/ : Participant age	0.7864	0.3187	2.468*

Table 4.5: Fixed effects in the regression model fit to the perceived ethnicity data. Symbols following the *t*-value indicate the associated *p*-value: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$, '.' $p < 0.1$

To begin the discussion of the effects in Table 4.5, note that unlike in the perceived age comparison the prediction of an effect of CONGRUENCY is not borne out. The variable CONGRUENCY shows neither a main effect nor any interaction effects.

Even though VOWEL VARIANT and SPEAKER GUISE don't interact as predicted, as seen in the absence of an effect of CONGRUENCY, they each show a main effect. The main effect of VOWEL VARIANT suggests that the participants responded considerably more slowly to the Anglo variants than to the African-American

variants of both /ɑ/ and /ʌ/. The main effect of SPEAKER GUISE suggests that the participants responded more slowly when they saw the photo of the female Anglo speaker than when they saw the photo of the female African-American speaker.

Next, there are two effects pertaining to the position of a trial in the experiment that resemble those seen in the perceived age data discussed above. The main effect of BLOCK shows that participants became faster the more blocks they had completed. The main effect of EXPERIMENT HALF shows that responses in the second half of the experiment were slower than in the first half.

Also parallel to the perceived age data, there is a main effect of PARTICIPANT AGE. Older participants responded more slowly than younger participants overall. There is also again no main effect of VOWEL. Neither /ɑ/ or /ʌ/ was responded to more quickly or more slowly than the other. However, VOWEL shows two interaction effects. The interaction of VOWEL and TRIAL IN BLOCK suggests that the vowel /ʌ/ elicited progressively slower responses than the vowel /ɑ/ in the course of each block. The interaction of VOWEL and PARTICIPANT

AGE suggests that older participants responded more slowly than younger participants to the vowel /ʌ/ than to the vowel /ɑ/.

When the participant's own ethnicity is added to the regression model, no significant main effects of ethnicity or interaction effects between ethnicity and the variable CONGRUENCY emerge. However, three of the four ethnicity groups show interaction effects with other variables.

First, adding the variable ANGLO to the regression model in Table 4.5 brings out one additional interaction effect. This interaction and the non-significant main effect of ANGLO are shown in Table 4.6.

Additional predictors of RT	Estimate	Std. error	*t*-value
Ethnicity Non-Anglo	-12.711	18.206	-0.698
Ethnicity Non-Anglo : Block	2.7828	0.6966	3.995***

Table 4.6: Changes in the regression model fit to the perceived ethnicity data when the variable ANGLO is entered

The addition of the interaction between ANGLO and BLOCK goes along with a change in the regression coefficient of the main effect of BLOCK from -3.50 to -5.49. Together, the main effect of BLOCK and the ANGLO : BLOCK interaction

suggest that while all participants' responses became faster across blocks, this effect was about twice as strong for the Anglo participants. Note that this is essentially the same effect found for Anglos in the analysis of the /eɪ/-/ɛ/ trials.

Adding the variable ASIAN to the regression model in Table 4.5 brings out one additional interaction effect. This interaction effect and the non-significant main effect of ASIAN are shown in Table 4.7.

Additional predictors of RT	Estimate	Std. error	*t*-value
Ethnicity Non-Asian	27.0757	19.0683	1.420
Ethnicity Non-Asian: Vowel /ʌ/	-22.0428	6.8537	-3.216**

Table 4.7: Changes in the regression model fit to the perceived ethnicity data when the variable ASIAN is added

The interaction between ASIAN and VOWEL suggests that non-Asian participants were faster, and Asian participants slower, in responding to the vowel /ʌ/ than to the vowel /ɑ/. This new interaction leaves unaffected the prior interaction between VOWEL and AGE. Also as previously, there is no main effect of VOWEL.

Finally, adding the variable AFRICAN-AMERICAN to the regression model in Table 4.5 brings out one additional interaction effect. This interaction effect and the non-significant main effect of AFRICAN-AMERICAN are shown in Table 4.8.

Additional predictors of RT	Estimate	Std. error	*t*-value
Ethnicity Non-African American	-19.4878	19.1576	-1.017
Ethnicity Non-African American : Trial in block	1.1939	0.4657	2.564*

Table 4.8: Changes in the regression model fit to the perceived ethnicity data when the variable AFRICAN-AMERICAN is added

The interaction between AFRICAN-AMERICAN and TRIAL IN BLOCK suggests that as more trials were completed per block, the African-American participants showed a greater decrease in their response time than the other participants. However, as in the original model in Table 4.5, no main effect of TRIAL IN BLOCK emerges.

4.3. Participant feedback

Participants' responses to the two post-task questions that formed part of the experiment showed a great deal of overlap with their comments in the open-ended debriefing after the experiment. Often participants repeated in the debriefing the same comments they had typed earlier in response to the two questions. Therefore, I will not discuss separately the answers to the written questions and the observations made in the debriefing. Instead, I first summarize and discuss all comments addressing the question of the difficulty of the task and then review all comments addressing the question about the speakers and their voices or accents.

Some participants made far more spontaneous comments than others, and different participants commented on different aspects of the experiment. This makes it difficult to evaluate tendencies in these data in quantitative terms. For this reason, I will not attempt to quantify these results. Instead, I will restrict the discussion to those trends which seem the most consistent in qualitative terms.

The participants overwhelmingly found the task easy to perform ("not that difficult," "pretty easy," "fairly easy"). Some explained that this was because the words were easy to understand and to distinguish. There was also widespread agreement that the speed at which the trials were presented was just right. In the debriefing, several participants explicitly said that they found the experiment interesting, but other comments suggest that for some participants the task became monotonous. One reported becoming "a little numbed by the back and forthness of the sounds," and another felt that the words "began to sound alike." One specific aspect that was pointed out as difficult was having to remember which word was mapped to which button. Recall that the mapping of the response alternatives to the left and the right button was displayed only at the beginning of each block. Some participants talked about having to force themselves to stay focused, and some mentioned losing track of the mapping at some point. Many participants mentioned which of the two trial types they found more difficult, the /eɪ/-/ɛ/ trials or the /ʌ/-/ɑ/ trials. They were about evenly split on this point. Some participants made specific comments about the /eɪ/ and /ɛ/ words. Those who found them easier explained that the final "d"

made it easy to decide between the words. Some of those who felt that the /eɪ/-/ɛ/ trials were the more difficult type noted that this was because the relevant words sometimes sounded very similar at first, i.e. at the beginning of the vowel, so that it was easy to make mistakes when rushing.

The participants' comments about the speakers' voices and accents included practically no concerns about the authenticity of the voices and speakers. In the debriefing, almost all participants were surprised to find out that the vowels were synthetically produced. Only one participant noted that the words sounded "electronically altered". This participant explained that he was sure about this because as a musician he regularly performs audio editing, including the editing of voice recordings. He reported hearing parts of words spliced together that did not belong together. Incidentally, this speaker's responses were excluded from the quantitative analysis because he had spent most of his adult life abroad, in accordance with the exclusion criteria discussed in Chapter 3. The only unrealistic aspect of the experiment that participants repeatedly commented on was hearing such a wide variety of different pronunciations from the same speakers. Some participants said they were

surprised to hear the speakers' pronunciation change back and forth, especially when hearing different pronunciations of the same word in sequence.

Not all participants reported hearing a particular accent in the speakers, but most did. Judging by their debriefing comments, at least for some participants the word "accent" was apparently misleading because to them it suggested either a stereotype, for example a Boston accent or "deep" Southern accent, or a foreign or other unfamiliar accent. Those who did perceive the speakers as having an accent almost all identified the accents as "Southern," "Texan," "Country," or characterized by "twang" or "drawl" ("they all sound like Texans", "definitely a Southern twang"). Some went on to explain that this did not surprise them as they were familiar with these accents, having grown up and lived in Houston. Others felt it was unusual to hear such Southern accents in Houston ("more of a drawl than most of the people I talk to from Houston", "just don't hear it every day", "a little more Country than I expected"). In a small number of cases, the Southern features were actually described as unpleasant.

> I just found some of the pronunciations annoying ... Michael was annoying. It didn't sound like a Houston accent. It was very Country sounding. Like "baaaay."

In discussing the speakers' accents the participants did not spontaneously mention race or ethnicity-affiliated speech. In fact, even when asked directly, the white female speaker was never described as sounding black. Instead, she was consistently described as sounding Southern, regardless of the participants' own ethnicity. Only one participant did not use "Southern" or one of the related terms ("Country," "twang," etc.) but described the "Southern" pronunciations as conveying particular attitudes. She noted that in some cases the speaker's attitude was more "casual," "slang," or "ghetto," and in other cases it was more "professional."

A surprising number of participants spontaneously commented on specific vowel qualities, especially during the debriefing. Some underscored their points by imitating the variant in question. In their comments, a striking correlation emerged. Male participants commented regularly on /ɛ/ and, to a lesser extent, /eɪ/, while female participants typically did not mention /eɪ/ and /ɛ/ at all and

instead talked about /ʌ/ and to a lesser extent /ɑ/. In other words, participants tended to comment on the pronunciations heard from speakers of their own gender.

Regarding the /ɛ/ and /eɪ/ variants, one participant gave a very precise rendition of a Southern raised /ɛ/, however not diphthongal [eə] as in the stimuli but triphthongal [æɪə]. Another participant correctly concluded that "the vowels in bay/bed and day/dead can be 'twanged' to sound more similar," again showing awareness of the relevant variants. Commenting on the same phenomenon, another participant went even further in suggesting that "the accent could have been played up more." In his written comments, he suggested "dayd" as a variant of *dead* that would have made *dead* and *day* sound even more similar. This participant's "dayd" presumably also refers to a triphthongal /ɛ/, i.e. to [æɪə]. It is interesting that most of the detailed comments were about /ɛ/ and not /eɪ/.

As for the /ɑ/-/ʌ/ trials, by far the most commented on vowel was /ʌ/, which was often performed as a hyper-raised and lengthened [ʊː]. In the written comments, raised /ʌ/ was described as "thick," "heavy," and "Country." On the

other hand, no one commented spontaneously on /ɑ/. Even where I was able to steer the participants' attention to the different variants of /ɑ/ in the debriefing, they typically reported no clear intuition about how the notion of "Southern" applies to /ɑ/ at all. Only one African-American participant gave a more centralized [a] as an example of a Black way of pronouncing /ɑ/.

When asked explicitly about differences between the two members of each pair of speakers, e.g. the two female speakers, some participants accurately reported that the two speakers sounded very similar, or in fact the same. This is accurate in that both speakers were using the same sociolinguistic variants, and both had the same *f0* parameters. One participant felt that the female voices were so similar that she thought that they had been switched at some point. Still, a greater number of participants reported that one of the two speakers in the pairs sounded at least somewhat different from the other. For the female speakers, there was a clear pattern in which the Anglo speaker, but not the African-American speaker, was heard as conspicuously Southern. This speaker was most frequently characterized as sounding "Country." The verbal comments often included a noticeable air of surprise or even what I felt bordered on

embarrassment in discussing this speaker's accent. For example, one participant recalled knowing "one or two girls" in her high school class who sounded like the speaker, clearly suggesting that that was not the norm but the exception for her peers. One participant commented that while it is conceivable that the speaker is from Katy – the ostensible home of this speaker – she would have to be from the Western (i.e., the more rural) part of Katy. By contrast, the African-American speaker received relatively few comments. The most specific ones came from the African-American participants, who merely noted that she sounded "normal", "familiar", or "easy to understand". One participant specifically mentioned that she had friends in Missouri City – the ostensible home of this speaker – who sounded exactly like her. While the comments on the female speakers show these fairly clear trends, for the younger and older male speakers there was no asymmetry. There were very few spontaneous comments singling one of the two out as different, for example as more Southern. And even when asked directly who, if any, sounded more Southern (or, more Texan), about equally many participants pointed to the younger and the older speaker.

In summary, the qualitative results emerging from the participants' comments show that the matched guise manipulation was successful. Participants appear to have had little if any doubt that the auditory stimuli were the voices of real people. Also, it appears that the experiment succeeded in creating different degrees of sociophonetic congruency, at least in the case of one of the female speakers. The African-American (or, as the participants put it, Southern) variants were noted in the Anglo speaker but were either not noticed or heard as "normal" in the African-American speaker. The results of the age comparison were not as clear-cut. The participants did not clearly report a lack of congruency in the case of the male speakers. Neither the younger nor the older speaker was clearly heard as unexpectedly Southern or non-Southern. Instead, both were heard as Southern.

Chapter 5

5. Discussion of the results

The results presented in the previous chapter appear contradictory. While one part of the experiment brought forth an effect of sociophonetic congruency, the other part of the experiment did not. In the trials involving the vowels /eɪ/ and /ɛ/ the participants responded more quickly, at least initially, when the social and linguistic information presented to them matched the dialect configuration found in Houston, and they responded more slowly to trials for which the opposite was the case. This result is consistent with the research hypothesis formulated in Section 1.4 and the specific predictions formulated in Section 2.4. On the other hand, in the part of the experiment in which listeners responded to trials involving the vowels /ɑ/ and /ʌ/ no effect of sociophonetic congruency was found. Complicating the results further, even in their responses to the /eɪ/

and /ɛ/ trials the participants displayed the predicted response bias only temporarily. As they completed more trials, they gradually came to display a bias in the opposite direction.

In the light of these results the research hypothesis that sociophonetic congruency influences speech perception is neither clearly supported nor can it be clearly rejected. Before conclusions can be drawn, two questions must be resolved.

1. Why did the listeners display a congruency bias in their responses to the /eɪ/-/ɛ/ trials but not in their responses to the /ɑ/-/ʌ/ trials?

2. Why did the effect of congruency in the case of the /eɪ/-/ɛ/ trials grow weaker and in fact reverse itself in the course of each block?

The purpose of this chapter is to provide a coherent explanation of the results as a whole before discussing their theoretical implications in Chapter 6. I will tackle the two questions in turn. Section 5.1 deals with the first issue. I argue

that while the lack of a congruency effect in the /ɑ/ and /ʌ/ trials may indicate a lack of sensitivity on the part of the listeners to sociophonetic congruency in their responses, it does so only in a limited way. The listener's behavior in this part must be interpreted relative to the task demands of the experiment. As I argue, the participants were able to pursue a response strategy which was not available to them in completing the /eɪ/ and /ɛ/ trials and which explains why a congruency effect was found in the latter but not in the former trial type. In Section 5.2, I turn to the second question. I argue that the change in the direction of the congruency effect does not contradict the research hypothesis for two reasons. First, the attenuation of the participants' congruency bias is a predictable response to the properties of the experimental stimuli. Prior findings on perceptual learning of dialects would in fact predict that the listeners "unlearn" their original assumptions. Second, more speculatively, the reversal of that bias can be accounted for as an effect of experiment-induced short-term perceptual learning in which the listeners gave more weight to incongruous information than to congruous information.

5.1 The lack of a congruency effect in the perceived ethnicity trials

It is not immediately obvious why a sociophonetic congruency effect was found in the /eɪ/ and /ɛ/ trials but not in the /ɑ/ and /ʌ/ trials. There is nothing in the production data discussed in Chapter 2 that makes the two cases of ethnicity-based variation less certain than the two test cases of age-based variation. All four effects emerge very clearly from the HUES wordlist data. Moreover, the participants' feedback summarized in Section 4.3 suggests that at least at the level of conscious awareness many participants experienced a lack of congruency in the female Anglo speaker's use of raised variant of /ʌ/. Their comments included a considerable amount of surprise at that combination. This makes it all the more surprising that there was no negative effect on response times, and that, instead, a response time effect was found in the case of the perceived age comparison where the participants' overt comments did not indicate strong surprise (see Section 4.3).

5.1.1 General listener-based dialect experience effects

In order to explain the contradictory results obtained in the two types of trials, I suggest considering first several other effects included in the regression models discussed in Chapter 4, i.e., effects not involving the variable CONGRUENCY. Several other significant effects are readily interpreted as reflecting general, long-term sociophonetic experience effects like those reported in the literature on dialect perception (e.g., Labov and Ash 1997, Clopper and Pisoni 2004b, Sumner & Samuel 2009). Such effects show that the more cumulative exposure a listener has had to a particular sociophonetic variant the more easily that variant will be recognized, regardless of the perceived identity of the speaker. For example, Labov and Ash (1997) showed that Chicago listeners have an advantage in recognizing local dialect variants such as fronted /ɑ/ even though their listeners were not given any specific information about who the speaker producing these variants was.

Applied to the /eɪ/ and /ɛ/ results, varying degrees of dialect exposure provide a straightforward explanation for the main effect of the variable VOWEL

VARIANT and its interactions with certain subgroups of participants. There were three effects. First, the Southern variants of both vowels, i.e., lowered /eɪ/ and raised /ɛ/, were each processed more slowly than the non-Southern variants by all participants, regardless of which photo the listeners saw. This was seen in the strong main effect of VOWEL VARIANT. The participants' greater difficulty in recognizing the Southern variants can be explained by the fact that the amount of non-Southern speech used by Anglos in Houston is larger overall than the amount of Southern speech. As discussed in Chapter 2, the dialect contact between linguistically Southern and non-Southern speakers in the Houston metropolitan area, as in other Texas metro areas (Thomas 1997), has created just such an imbalance historically.

Second, the interaction between VOWEL VARIANT and PARTICIPANT AGE in the /eɪ/ and /ɛ/ trials lends itself to an experience-based explanation as well. In fact, it supports the interpretation given in the previous paragraph for the main effect of VOWEL VARIANT. In the context of the decline of Southern dialect features in Houston, older speakers will have experienced more Southern speech in their lifetime than younger speakers. They may, in fact, use more Southern variants

themselves. Therefore, it makes sense that they recognize words containing the Southern variants more quickly than younger listeners.

Third, general dialect experience can also be used to account for the fact that the group of Hispanic participants responded more slowly to the Southern vowels variants than all other groups, as seen in the interaction between VOWEL VARIANT and HISPANIC. It is conceivable that this group, which includes several Spanish-English bilinguals, has had less experience with the phonetic details of Southern Anglo speech because they have been exposed to less Anglo speech overall. In a somewhat parallel scenario, Preston (2005) found that African-American listeners in Michigan were less accurate in recognizing advanced variants of a sound change occurring in the Anglo mainstream variety. There is also anecdotal evidence for this interpretation in the participants' feedback. In the debriefing, one Hispanic participant noted that although she didn't notice anything unusual in the Anglo speakers' speech she wasn't sure because, as she said, she interacts mostly with Spanish speakers in day-to-day life.

Overall then, an account in terms of long-term dialect experience readily captures all effects of VOWEL VARIANT observed for the /eɪ/-/ɛ/ trials. In light of

this, it is striking that in the perceived ethnicity part of the experiment listeners responded more quickly to the African-American variants, both in the case of /ɑ/ and in the case of /ʌ/. This finding clearly clashes with an experience-based account. While it is conceivable that some of the participants may have had more cumulative exposure to African-American speech in their lifetime, specifically the African-American participants, it is highly unlikely that *all* participants had a greater amount of experience with fronted /ɑ/ and raised /ʌ/ than with the respective Anglo variants. Recall that instead of an interaction between the variables VOWEL VARIANT and AFRICAN-AMERICAN, a main effect of VOWEL VARIANT was found. Thus, an explanation in terms of dialect exposure fails to account for the general RT advantage of the African-American variants. There has to be an alternative explanation for the unexpected direction of the effect of VOWEL VARIANT in the /ɑ/ and /ʌ/ trials.

5.1.2 Differences in the degree of phonetic distinctiveness

I suggest that the unexpected advantage of the African-American variants of /ɑ/ and /ʌ/ can be explained with reference to an asymmetry in the phonetic properties of the stimuli heard in both parts of the experiment. One difference between the two trial types is the general degree of phonetic distinctiveness of the two vowels involved in each. To illustrate this difference, Figure 5.1 shows the acoustic quality of all three variants of /eɪ/ and /ɛ/ in terms of the frequencies of the first two formants at a time point one third into the vowel. To ensure greater readability, the F1/F2 coordinates of the two /eɪ/ words (*bay* and *day*) and the F1/F2 coordinates of the two /ɛ/ words (*bed* and *dead*) were averaged in Figure 5.1. Figure 5.2. shows the acoustic quality of the three variants of /ɑ/ and /ʌ/ at the same time point.

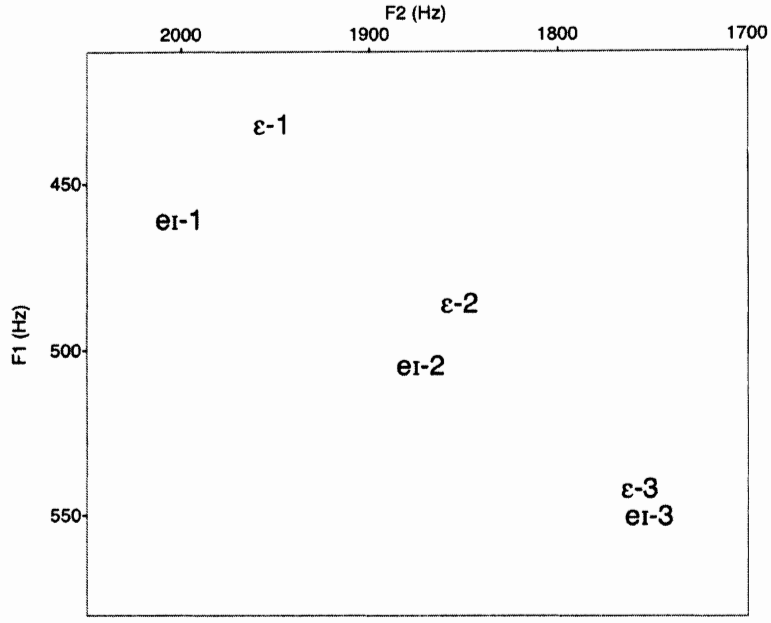

Figure 5.1: Acoustic quality of variants 1 and 3 of /eɪ/ and /ɛ/ at a time point one third into the vowel

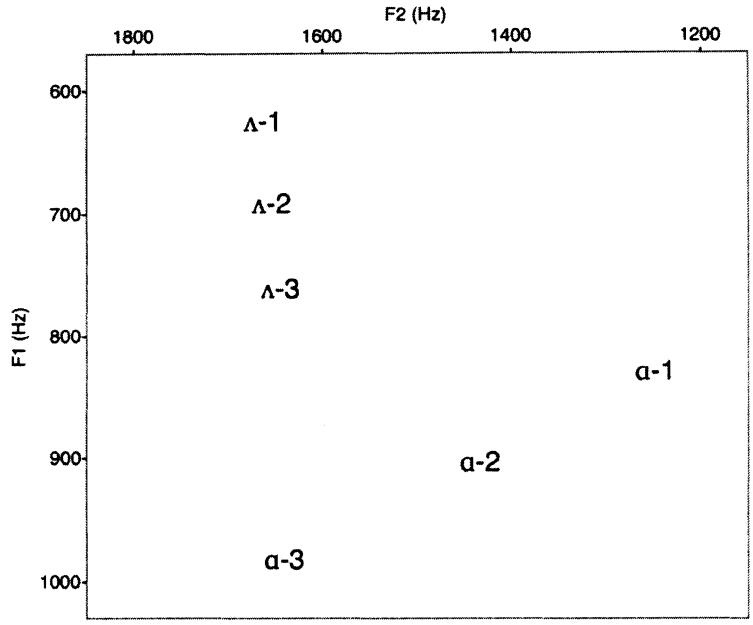

Figure 5.2: Acoustic quality of variants 1 and 3 of /ɑ/ and /ʌ/ at a time point one third into the vowel

The position of the vowel variants in F1/F2 space shows that at the time point represented in Figure 5.1 and Figure 5.2 the general degree of distinctiveness of the /eɪ/ and /ɛ/ stimuli relative to each other is much smaller than the general degree of distinctiveness of the /ʌ/ and /ɑ/ stimuli. For example, the quality of Southern /eɪ/ and non-Southern /ɛ/ are practically identical at this point. By contrast, the variants of /ɑ/ and /ʌ/ occupy distinct areas in F1/F2 space. They partially overlap in F2 but they do not overlap at all in F1.

What are the likely perceptual consequences of this asymmetry? F1 and F2 are initially of very limited value in identifying a stimulus word as, for example, *day* or *dead*. It appears that although the /eɪ/ and /ɛ/ variants are not globally ambiguous they are temporarily ambiguous at an early stage of word recognition, and therefore clearly more confusable than the variants of /ɑ/ and /ʌ/. This initial similarity of /eɪ/ and /ɛ/ was in fact pointed out by some of the participants in their comments discussed in Section 4.3. As one of them put it, when /eɪ/ and /ɛ/ are "twanged", they become more similar. And, as another pointed out, this made it easy to mistake one vowel for the other when trying to respond quickly. By contrast, formant frequencies, especially F1, provide a very effective cue to identifying /ɑ/ and /ʌ/ from the beginning. As can be seen in Figure 5.2, within the range of variation heard in the experiment any stimulus with an F1 higher than 800 Hertz was unambiguously identifiable as the vowel /ɑ/, and any stimulus with an F1 lower than 800 Hz was unambiguously identifiable as the vowel /ʌ/.

5.1.3 Task demands

What does this difference in the degree of phonetic distinctiveness of the stimulus vowels mean for the likely response strategies of the listeners? Here we need to reconsider the nature of the task itself. The fastest way for the participants to identify the stimulus word as, for example, *day* or *dead*, was for them to identify the vowel which distinguished the two response alternatives. In the case of /ɑ/ and /ʌ/, the vowel was in fact the only way to distinguish them. Because the response set contained only two words, their task was essentially that of discriminating between the vowels. In such a situation, they would have benefited from paying particular attention to those acoustic phonetic properties which distinguish one vowel from the other. The more such cues there are, the easier it is to come to a decision quickly.

In the case of /ɑ/ and /ʌ/, participants were able to base their responses on a small set of phonetic cues, notably the frequency of the first formant. That is because, as discussed above, F1 unambiguously distinguishes all /ɑ/ and /ʌ/ variants. Such an F1-based strategy provides a straightforward explanation for

why the fronted /ɑ/ and the raised /ʌ/ showed a processing advantage over the backed /ɑ/ and the non-raised /ʌ/. The former two were responded to more quickly because they show a greater difference in F1 relative to the acoustic space occupied by the other vowel. In other words, the raised /ʌ/ is the least /ɑ/-like of the variants of /ʌ/, and the centralized /ɑ/ is the least /ʌ/-like of the variants of /ɑ/. When following such a strategy, it may be seen as coincidental that these variants are also the ones associated with African-American English in Houston. It appears that the unexpected effect of VOWEL VARIANT, apparently pointing to an advantage of the African-American variants, is not an effect of dialect experience but an effect of the experimental design, specifically the small number of response alternatives and their phonetic properties.

Compare this with the task demands of the perceived age trials. Here, the temporary phonetic similarity of the /eɪ/ and /ɛ/ variants early in the vowel made it impossible to rely exclusively on an acoustic phonetic strategy. In this situation, non-phonetic information becomes potentially relevant because it helps predict which vowel is more likely to be heard. Such information was, of course, made available through the photo manipulation. It is such a strategy,

then, that allowed the predicted effect of sociophonetic congruency to emerge in the /eɪ/-/ɛ/ trials but not in the trials involving /ʌ/ and /ɑ/. As I discuss further in Chapter 6, this explanation of the congruency effect is, of course, very similar to the explanation of similar effects in designs based on global phonological ambiguity given in Chapter 1.

In summary, the design of the experimental task can be seen as having influenced the likelihood of observing the predicted effect of sociophonetic congruency. The fact that there were only two response alternatives allowed the participants to pay exclusive attention to a narrow range of acoustic phonetic cues, viz. those which distinguish only the two vowels in question. This created an incentive for them to ignore the social information in one part of the experiment because the most efficient response strategy was a purely phonetic one. The apparently contradictory finding of a congruency effect in one part of the experiment but not in the other is due to the fact that this response strategy was available only in one part. Where the participants could use the phonetically-driven strategy there was no measurable effect of sociophonetic congruency. Where the participants could not use this strategy, an effect of

congruency was found. It emerged here because the listeners took a broader spectrum of cues into account, including the social cues provided by the picture of the speaker. This interim conclusion raises the question which of the two trial types speaks more directly to the research hypothesis. I will take up this question in the general discussion in Chapter 6.

5.2 The reversal of the congruency effect in the perceived age trials

The second major challenge to the research hypothesis comes from the finding that even in the part of the experiment in which participants displayed the predicted response bias they did so only temporarily. In fact, the more /eɪ/ and /ɛ/ trials they completed per block the more they came to respond in a way that is exactly opposite of the predicted direction. This was seen in the interaction of the variables CONGRUENCY and TRIAL IN BLOCK discussed in Section 4.2. In approaching this question, I suggest distinguishing between two separate phenomena. The first is the finding that the participants' response bias grew

weaker in the course of completing more trials. The second is the finding that their response bias actually reversed itself.

The gradual attenuation of the congruency effect which occurred roughly within the first half of each block of trials (see Figure 4.1) does not call into question the validity of the congruency bias. It merely suggests that the participants quickly adapted their response strategy by replacing their original biases with new biases learned from the experimental stimuli. There are clear precedents for such an effect in the experimental literature on dialect perception. For example, Clarke & Garrett (2004) provide evidence that native English listeners rapidly adapt their perception when listening to foreign-accented speech. Maye, Aslin & Tanenhaus (2008) showed that listeners are able to spontaneously and quickly adapt to a novel accent similar to a regional dialect of English. Kraljic & Samuel (2007) provide evidence that listeners are even able to quickly adjust their perception to multiple speakers displaying different types of pronunciation variation. These studies show that listeners have a capacity for rapid perceptual adjustment to specific phonetic features that are encountered.

Applied to the current results, it is not difficult to see how rapid perceptual adjustment to the stimulus speakers would lead to the disappearance of the congruency effect. Recall that the younger and the older Anglo speaker were each heard using both the Southern and the non-Southern variants of /eɪ/ and /ɛ/ to equal degrees. Thus, while their "dialect" was heterogeneous in that it was neither consistently Southern nor consistently non-Southern, it was consistent in that both variants occurred at exactly the same rate in each speaker. This allowed the participants to learn that both speakers were equally likely to use both variants and, in response, abandon their prior biases and anticipate hearing all variants equally often. As a result, they would no longer respond more slowly to one type than to the other.

In the following, I will continue to refer to the process in which the participants adapted their responses to the properties of the experimental stimuli as "perceptual learning." I am doing so even though it appears that their "learning" was temporary. Recall that the effect did not carry over from one block of trials to the next. The results of the regression analysis in Chapter 4 showed no interaction between the variables CONGRUENCY and BLOCK. Thus, the

participants appear to have "forgotten" what the two male speakers sounded like between the end of one block and the start of the next block featuring one of them. Under a strict definition of perceptual learning (e.g., Goldstone 1998), such a temporary effect does not qualify as perceptual learning. However, I will use that term here because the effect appears to be otherwise in line with what studies of more permanent perceptual learning have found.

It may be helpful to point out why other studies which have demonstrated effects of sociophonetic knowledge on speech perception (e.g., Drager 2005, 2011; Hay, Warren & Drager 2006) did not find a similar learning effect. In these studies the listeners were not given positive feedback regarding the stimulus speakers' degree of /æ/-raising or /iə/-/eə/ merger, respectively. Rather, the listeners had to decide for themselves what phonemic category each phonetic token that was heard belonged to. They were not in a position to infer each speaker's dialect from the information in the trials as they could not be sure whether any of their own responses were correct. In the current experiment, on the other hand, the extremely low error rate shows that at the end of the trial participants knew which word they had heard, and therefore also

which vowel variant was used by each of the stimulus speakers. This allowed them to build up speaker-specific representations of each stimulus speaker's "dialect."

Having motivated the gradual attrition of the congruency effect, the next challenge is to explain the listeners' apparent overcompensation. Why did their response bias not disappear but instead reverse itself? This finding is inconsistent with the idea that the participants merely adapted to the likelihood with which each variant was heard. It suggests, instead, that they arrived at a new response bias. What led them to this new bias? To my knowledge, there is no precedent of such an effect in the experimental literature on dialect perception. However, I believe that a plausible explanation comes from the notion of sociophonetic congruency itself, when combined with the idea of differential attention. This post-hoc explanation is admittedly more speculative and will require further research to be substantiated.

Note that the perceptual learning process described above presupposes that both the Southern and the non-Southern vowel variants were given equal weight in the listeners' perceptual adjustment to the two stimulus speakers'

"dialects." However, it is not clear that this is how perceptual learning of dialects works. If it is true, as was predicted here, that listeners respond differently to sociophonetically congruous and incongruous speech, it is conceivable that congruous information and incongruous information are also treated differently in perceptual learning. Specifically, it is conceivable that incongruous information is given greater weight and thereby has a larger impact on the participant's emerging model of each speaker.

Why should incongruous information be weighed more heavily? This would be expected if perceptual learning is mediated by the degree of attention that listeners pay to different types of information. It's possible that the incongruous trials, because they contradict the listeners' prior assumptions, drew a greater amount of attention at first. If additional attention leads to greater learning, incongruous trials contribute more to the emerging model of the perceived speaker's dialect. This explains why the two speakers came to be thought of, apparently erroneously, as more Southern or less Southern than warranted by the actual quality of the stimuli. On the basis of this new representation, in which, for example, the younger speaker is expected to use

the Southern variants more than the non-Southern variants, in the later trials the Southern variants were then responded to more quickly and the non-Southern variants more slowly. This explains the apparent reversal of the participants' response bias.

This interpretation of the reversal of the congruency bias does not contradict but in fact supportsthe claim that the listeners responded differently to congruous and incongruous information. However, it crucially involves the idea that differential attention to congruous and incongruous information caused the effect. I will return to this point in Section 6.2.

Chapter 6

6. General discussion and conclusions

In this final chapter I discuss the findings of this dissertation further in order to arrive at a general conclusion and in order to spell out their theoretical implications. Section 6.1. summarizes the main results and the hypothesis which gave rise to them. In Section 6.2, I revisit the theoretical debate over the question under what conditions social information is accessed in speech perception. In Section 6.3. I discuss the implications of the current results for examplar-based models of sociophonetic knowledge and learning.

6.1. Summary of the main findings

The goal of this dissertation was to clarify the role which social information about a speaker plays in the phonetic perception of his or her speech. I have used the term *sociophonetic knowledge* to refer to implicit assumptions which

language users in a particular speech community have about how members of different groups of speakers in that community produce speech sounds. Previous research has repeatedly demonstrated that sociophonetic knowledge can influence speech perception (Strand & Johnson 1996; Niedzielski 1997, 1999; Drager 2005, 2011; Hay, Warren and Drager 2006) and that the body of knowledge which listeners access in this process is large and fine-grained (Hay, Warren and Drager 2006). However, as argued in Chapter 1, it is not clear exactly under what circumstances sociophonetic knowledge is put to use. This is because much prior work relied on a particular line of evidence, the variable resolution of global phonological ambiguity. Arguably, effects of sociophonetic knowledge which emerge in such a task are not wholly representative of speech perception. That is because global ambiguity resolution requires additional processing effort and, as argued by Luce, McLennan & Charles-Luce (2003), listeners access social-indexical information, or what I have called sociophonetic knowledge, only if word recognition is effortful and slow.

The specific question asked in this dissertation was therefore: Does sociophonetic knowledge inform speech perception where listeners are not faced

with global lexical ambiguity? Strand's (2000) finding of an effect of gender typicality in a shadowing task, although not entirely conclusive from a sociolinguistic point of view, suggested that this may be the case. This hypothesis was tested experimentally in the context of sociophonetic variation in the Houston metropolitan area. The experimental manipulation resulted in the predicted effect of sociophonetic knowledge in one part of the experiment, the one where the listeners heard words containing the vowels /eɪ/ and /ɛ/, but not in the part in which they heard words containing the vowels /ɑ/ and /ʌ/. In Chapter 5 I offered an explanation for this apparent contradiction. I argued that the lack of an effect in one part does not invalidate the finding of the predicted effect in the other part because the participants' responses in both parts can be given a unified explanation with reference to the task demands of the experiment. This explanation required a reconsideration of the role of ambiguity in bringing out effects of sociophonetic knowledge. In Chapter 5 I also discussed the unanticipated result that the listeners appeared to progressively unlearn their original congruency bias in the course of completing more trials per block and come to display a response bias in the opposite direction. I offered a post-

hoc explanation for this in terms of differential attention to congruous and incongruous information in perceptual learning.

6.1 The role of sociophonetic knowledge revisited

As argued in Chapter 5, the non-identical task demands of the different parts of the experiment help explain the apparently contradictory results. The crucial difference was the degree of acoustic phonetic distinctiveness of the two vowels which distinguished the two response alternatives in each case. As there were only two alternatives, whenever the phonetic contrast between the vowels was sufficient to come to a decision on the basis of this information alone the participants did not access their sociophonetic knowledge. As a result, no effect of social information was found. This was the case for the vowels /ɑ/ and /ʌ/. On the other hand, where the phonetic contrast between the vowels was at least temporarily reduced, as in the early part of the vowels /eɪ/ and /ɛ/, listeners took the social information into account and an effect of sociophonetic knowledge emerged. This interpretation of the results leads to the generalization

that the role of social information in speech perception is correlated with the degree of linguistic ambiguity listeners are confronted with, even if that ambiguity is temporary. Thus, as argued in Chapter 1 for the effect of global lexical ambiguity, listeners use social information if the linguistic information available to them does not allow one of several lexical candidates to be selected. The current results show that they do so even where the social information is not strictly required to disambiguate a word because the ambiguity is temporary. That is, the listeners had the option of ignoring the social information and wait, as it were, for the linguistic ambiguity to be resolved by itself.

Two general conclusions can drawn from these results. First, speech perception may be influenced by sociophonetic knowledge even where there is no global but only temporary ambiguity. Second, where there is no linguistic ambiguity whatsoever, listeners are not measurably affected by sociophonetic knowledge.

The first conclusion, that temporary ambiguity is a sufficient condition for listeners to access sociophonetic knowledge, entails that the influence of social

information on speech perception is considerably more widespread than previous studies were able to demonstrate. To estimate the generality of the effect, it is helpful to compare the nature of the linguistic ambiguity created in the present task with the nature of the ambiguity involved in previous sociophonetic speech perception studies. In a phonetic continuum categorization task (Strand & Johnson 1996; Johnson, Strand & D'Imperio 1999; Drager 2005, 2011) where participants identify the stimuli as one of two lexical items distinguished only by the value of the relevant sociolinguistic variable, e.g. the quality of the vowel in the word *bad* or *bed* (Drager 2005, 2011), the category boundary shift brought about by seeing the image of the speaker is observable especially in the intermediate tokens, i.e., those which are closest to the category boundary and thus the most ambiguous ones. This shows that this type of task is based specifically on listeners' responses to extreme, in fact possibly complete linguistic ambiguity. The listeners are practically pushed to make inferences from any non-linguistic information that is available to them because the linguistic information itself is of little or no value in completing the task. On the other hand, in a shadowing task like that used by Strand (2000) the words

are unambiguously identifiable in all trials. Ambiguity resolution is required here only in a more limited sense. In order to be identified, each word had to be distinguished from all initially identical words in the English lexicon. The process of word recognition starts as soon as the speech signal is perceived and can be conceptualized as the stepwise elimination of cohorts of competitors (Marslen-Wilson & Tyler 1980). Therefore, before the word's uniqueness point is reached, there is ambiguity between the target word and the current competitors. However, crucially, this type of ambiguity is temporary. Coming back to the current results, the type of ambiguity resolution that the participants performed here is clearly more similar to the type of ambiguity faced by Strand's (2000) listeners. There was a time period in which the available linguistic information was incomplete so that inferences drawn from non-linguistic aspects of the situation would be of potential value but the participants were not in a position where they *had* to make use of sociophonetic knowledge because the ambiguity was temporary.

Thus, temporary ambiguity resolution, unlike global ambiguity resolution, is a fairly common occurrence in word recognition. After all, practically any

word must be initially distinguished from a pool of competitors before it is recognized. Therefore, while social considerations may not *invariably* make a difference in speech perception, as seen when there are no competitors at all and thus no ambiguity to be resolved, it appears that the situations in which listeners would potentially benefit from sociophonetic knowledge occur quite frequently. The present experiment therefore demonstrates that sociophonetic knowledge has a considerably wider role in speech perception than previous sociophonetic studies were able to demonstrate.

How does this conclusion relate to Luce, McLennan & Charles-Luce's (2003) time course hypothesis? Recall that according to Luce et al.'s hypothesis, word recognition is mediated by indexical knowledge only when the time it takes to process a word exceeds a certain duration. When processing is rapid there is no time for indexical effects to emerge. Evaluating the present results from this perspective is difficult because the task which the participants in the current experiment performed was different from the tasks which gave rise to Luce et al.' hypothesis. The present results were based on a two-alternative forced choice word identification task. As discussed in Chapter 1, in Luce &

Lyons' (1998) long-term repetition priming study the participants did not show an effect of indexical variability in a lexical decision task. The authors did, however, find an indexical variability effect in a task in which the listeners were asked to decide whether the word was "old" or "new," which required more time to decide. Luce et al. argue that the first task represents relatively easy processing because the participants merely have to access the lexical item in the mental lexicon, whereas the second type of task incurred more processing effort. This difference was reflected in the participants' response times.

Given the disparity between the task and data types involved in the current study and Luce et al.'s study, no conclusive answer can be given here to the question whether the current results support or contradict Luce et al.'s time course hypothesis. However, there are reasons to believe that the degree of difficulty of the current task was so low as to fall under what Luce et al. consider rapid processing. The two-alternative forced choice design of the current experiment can be seen as having created a particularly easy task in both trial types. The participants' response was facilitated by the fact that, unlike in a lexical decision task, the response alternatives were explicitly presented. In

addition, there were only two response alternatives so that the number of competitors was already reduced to one. Clopper, Pisoni & Tierney (2006) have shown that closed set tests of word recognition, as in the current experiment, lead to greater rates of correct word recognition the smaller the number of response alternatives is. It is only once the number of alternatives is relatively large that the effects seen in closed set tests resemble those seen in open set tests. Thus, a two-alternative forced choice design would seem to aid processing considerably. This should have resulted in faster processing than if participants had identified the same words from an open set.

If it is the case that the responses to the /eɪ/ and /ɛ/ trials involved rapid processing in the sense of Luce et al., the current results constitute counterevidence to their time course hypothesis. Although this possibility is speculative, it raises the question why it should be that in their experiments indexical specificity effects were mediated by the speed of processing but not in the current experiment. I believe that such a difference would be due to the differing nature of the indexicality effects involved. The type of indexicality studied by Luce and colleagues is, after all, different in several respects from

what I have called sociophonetic knowledge in this dissertation. The main difference is that their effects were experimentally induced, while the current experiment relied on biases acquired through long-term experience with language variation in a speech community. Because of this, the total amount of experience which the social indexicality effects at issue in the present experiment rest on is considerably larger. If more cumulative experience leads to stronger indexicality effects, the sociophonetic knowledge as defined here would have a more powerful influence on speech perception than talker variability learned in a repetition priming design.

As note above, this conclusion is obviously speculative because it is not completely clear that the responses to the /ei/ and /ɛ/ trials in the current experiment count as rapid responses. To decide whether they do, one would have to replicate Luce & Lyons' (1996) and McLennan & Luce's (2005) experimental designs using the stimuli created for the current experiment. That is, one would have to measure the response time to these stimuli in a lexical decision task, as well as a shadowing task, and include many additional stimuli spoken by same and other voices. As I discussed in Chapter 3, however, the use

of speech synthesis, especially the time required to produce highly realistic and precise phonetic stimuli conflicts with the demands of constructing such a task.

Regarding future experimental designs, the contrast between two sources of social indexicality discussed here – variation learned in the context of an experiment and variation learned in a community of speakers – demonstrates the usefulness of incorporating existing community variation into the design of experimental studies of indexicality effects on spoken word recognition. The advantage of drawing on participants' prior sociophonetic knowledge is that the types of linguistic variation which are studied are ones which are known to occur in society. This makes conclusions regarding indexical variability effects more realistic. Of course, drawing on existing community variation is more difficult in practical terms as it presupposes the availability of production survey data. However, I believe that the current study shows that doing so can be quite revealing for both sociolinguists and researchers interested in spoken language processing. Given the amount of survey work continuously conducted by sociolinguists and dialectologists and the degree of phonetic detail involved in many modern surveys (see, e.g., Labov, Ash & Boberg 2006) there appears to be

a large potential here for formulating precise hypotheses about speech perception.

6.3 Implications for exemplar models of sociophonetic knowledge

Several of the authors of the sociophonetic speech perception studies cited throughout this dissertation (Hay, Warren & Drager 2006; Hay, Nolan & Drager 2006; Drager 2005, 2011) have drawn on the framework of exemplar theory in interpreting their results. The assumptions of exemplar theory, originally developed in cognitive psychology to account for non-linguistic categorization (see Goldinger's 1997 review), lend themselves well to explaining the interconnectedness of social and linguistic information in speech processing. Therefore, I conclude this dissertation by discussing the findings of the current study with respect to their implications for an exemplar-based view of sociophonetic knowledge.

6.3.1 Exemplar-based models of sociophonetic knowledge

In exemplar models of phonological and lexical representation (Johnson 1997, 2006; Goldinger 1998; Pierrehumbert 2001) speech production and perception are conceptualized as involving the activation of large numbers of separately stored instances, or exemplars, of experienced speech events. In principle, every encountered utterance is thought to leave a unique memory trace. These exemplars are reactivated in subsequent language use. Lexical or phonological categories are understood as consisting of distributions of remembered exemplars so that activating a linguistic category means to activate the associated distribution of exemplars, or "exemplar cloud."

One attractive feature of exemplar models is the way in which they model the connection between general linguistic categories such as words or phonemes and individual usage events. Given that a general linguistic category such as a word or phoneme is directly instantiated by remembered usage events, newly stored exemplars immediately impact the representation of that category. Within linguistics, this aspect of exemplar models has proven particularly useful in

explaining frequency effects such as frequency-correlated phonological reduction (Bybee 2001; but see Labov 2006 for a critical assessment of the role of frequency in sound change). For phoneticians, and especially sociophoneticians, exemplar models are attractive because another key property of exemplars is that they preserve, in principle, all the phonetic detail included in the original experience associated with hearing or producing speech. This includes, for example, dialectal features present in a particular speaker. While such detailed features may be tangential to the linguistic information conveyed by an utterance, they are crucial when it comes to explaining memory for voices (Johnson & Mullennix 1997) and, as in the current context, memory for sociolinguistic variation.

Moreover, exemplar representations are not limited to acoustic information. They are conceptualized as holistic representations of experience. This means that they may include, for example, aspects of the context in which an utterance was encountered, such as the social identity of the speaker. Hay, Nolan & Drager (2006) refer to the process whereby phonetic information is stored together with information about the speaker as *social indexing.*

Phonetically detailed, socially indexed memories of speech constitute what I have called sociophonetic knowledge, or knowledge of how phonetic variation is distributed across societal groups.

The first conclusion of this dissertation, that sociophonetic knowledge affects speech perception even under what can be considered easy processing conditions, is fully compatible with the predictions of exemplar models. From an exemplar perspective, the congruency effect observed in the participants' responses to the words containing /eɪ/ and /ɛ/ can be explained as follows. The conceptualization of the two male voices as the voice of either a younger or an older speaker, due to the experimentally created speaker guise, caused the participants to activate exemplars of the relevant words spoken by Houston Anglo speakers of these age groups. Therefore, when conceptualizing the speaker as an older Anglo, the participants activated more exemplars of the words spoken with Southern vowel variants than when conceptualizing the speaker as a younger Anglo. As a result of having been pre-activated, the exemplar cloud associated with words containing Southern /eɪ/ and /ɛ/ variants as a whole receives full activation faster when an auditory stimulus containing those

variants is perceived than when the auditory stimulus contains a non-Southern variant. The current study can therefore also be understood as a test of the predictions of exemplar-based models of sociophonetic knowledge.

However, two findings in this dissertation are not easily accounted for by exemplar theory and therefore merit further discussion. The first is the finding that the participants were not measurably affected by sociophonetic congruency in one part of the experiment. It is not clear how their response strategy in the /ɑ/ and /ʌ/ trials can be given an exemplar-based explanation. The other is the finding that the participants failed to learn that the speakers were using Southern and non-Southern variants to equal degrees and, instead, appear to have perceptually overemphasized the import of some variants. These two issues are discussed in the two final sections below.

6.3.2 Selective activation and deactivation of exemplars

The first conclusion drawn from the experiment reported here is that speech perception is more widely affected by sociophonetic knowledge than previous

studies were able to demonstrate. However, the second conclusion is that in the absence of linguistic ambiguity listeners may respond in a way that is not measurably affected by sociophonetic knowledge at all. How can an exemplar account explain this? In Chapter 5, I described the perceptual strategy which the listeners used in this part of the experiment as one in which they paid exclusive attention to those features which distinguish /ɑ/ and /ʌ/, regardless of what dialectal variant of /ɑ/ and /ʌ/ was heard. This strategy was made possible by the fact that the range of dialectal variation in these two vowels does not include any overlap (see Chapter 2). Speaking in terms of exemplars, in the trials featuring the African-American speaker, the listeners did not selectively activate African-American exemplars, and in the trials featuring the Anglo speaker they did not selectively activate Anglo exemplars. If they had done so, this should have resulted in a congruency effect, which, however, was not observed. Apparently, the participants were activating *all* experienced exemplars of /ɑ/ and /ʌ/, or perhaps all 18-year old female exemplars, equally. This means that they must have been able to keep from automatically activating only some exemplars and not others. The fact that the participants were able to respond in

this way shows that there must be a mechanism by which participants can selectively ignore, or "tune out," sociophonetic knowledge where this knowledge is not useful.

This conclusion may be seen as conflicting with earlier results of sociophonetic experiments, for example the results of the vowel matching tasks of Niedzielski (1997, 1999), Hay, Nolan & Drager (2006) and Hay and Drager (2010). As described in Chapter 1, in these experiments the listeners were asked to perform a metalinguistic task in which they matched the quality of a vowel heard in a sentence context with tokens from a synthetic vowel continuum. The repeated finding in each of these studies was that the listeners' perception was biased by social cues presented to them. In one of Hay, Nolan & Drager's (2006) experimental conditions, the participants matched the vowels they had heard to more Australian sounding tokens than a control group even though the speaker was in fact a New Zealander. They apparently did so because the instructions suggested that the speaker was from Australia. However, they later reported that they did not believe the speaker to be Australian. This suggests that they were unable to "tune out" their sociophonetic knowledge and respond in a way that

was completely unbiased by it. More dramatically, Hay and Drager's (2010) follow-up experiment showed that effect found by Hay, Nolan & Drager (2006) was present even where the participants were given no overt cue to the dialect of the speaker at all. In their experiment, the participants were influenced merely by seeing either a stuffed toy kangaroo or a stuffed toy kiwi, iconic animals associated with Australia and New Zealand, respectively. This means that the participants should have been in an even more advantageous position to perform the matching task accurately, i.e., without selectively activating the Australian or New Zealand exemplars. But they still did not do so.

In the light of these contradictory findings, it is not clear whether an exemplar model of sociophonetic knowledge should allow the intentional, selective activation and de-activation of particular exemplar distributions or not. The present results appear to call for such a mechanism, while the results of the other studies discussed above don't appear to allow it. A possible solution to this problem may be that selective activation is indeed a possibility, but only where a listener knows *what* to ignore. Note that the participants' task in the current study differs from the vowel matching studies in one way. In the current

experiment the listeners knew that they would hear specific words containing /ɑ/ and /ʌ/. Thus, they were able to anticipate with a fair amount of precision the phonetic range of the variants that they would hear. This in turn allowed them to follow a strategy in which they anticipated hearing exactly two types of stimuli, more /ɑ/-like or more /ʌ/-like ones. They could do this by deactivating all exemplars of other words and activating all exemplars of the relevant /ɑ/ and /ʌ/ words equally. The participants in the vowel matching studies, however, had no equally clear motivation for responding in a neutral way. Hay, Nolan and Drager's (2006) participants reported not believing that the speaker they heard was Australian. However, they also did not have an incentive to respond as if they knew this with certainty. That is, even though they may have suspected being duped into believing that someone from New Zealand was Australian, they also were not told explicitly to ignore the erroneous instruction. Therefore, their responses can be seen as reflecting uncertainty about the veracity of the speaker's ostensible background, but not certainty that the instructions they were given were false. Had the instructions been to ignore the information that the speaker was Australian, they might have responded in a way that showed no

influence of their knowledge of Australian vowel variants at all. The participants in Hay & Drager's (2010) follow-up experiment were not consciously aware of being manipulated at all. Therefore, they were clearly also not in a position to respond in a way that deliberately ignores the manipulation.

In conclusion, exemplar-based models of sociophonetic knowledge need to and can include a mechanism for the selective activation and deactivation of specific exemplar distributions, as required by the current results, without contradicting prior findings. It should be noted, however, that such a mechanism may not be particularly useful in actual speech perception. After all, hypotheses about a given speaker's likely language variety are generally useful. There seem to be few if any real-life contexts in which expecting a novel speaker, let alone a known speaker, to use *any* variety would lead to faster word recognition. As discussed in Section 6.2 in connection with Clopper, Pisoni and Tierney (2006)'s findings regarding closed-set effects on word recognition, the processing conditions which caused the participants to pursue such a strategy are not very realistic because in real life the number of competing lexical alternatives is almost invariably much larger.

6.3.3 The role of attention in exemplar-based learning

The other challenge to exemplar-based models of sociophonetic knowledge posed by the current results is the finding, discussed in Section 5.2, that the participants in the current study apparently failed to learn that the two male stimulus speakers were both using exactly equal amounts of Southern and non-Southern variants of the vowels /eɪ/ and /ɛ/. If their perceptual learning process was driven simply by the rate of occurrence of each variant in the experiment, they would have learned to expect them to occur equally in each speaker and the congruency effect should have become attenuated and eventually disappeared. However, the results show that the participants effectively reversed their expectations. Their response times in the second half of the 24 trials per block suggest that they came to anticipate, for example, Southern vowel variants rather than non-Southern ones in the younger speaker's speech. In Section 5.2, I suggested that this unexpected learning effect may have been driven by selective

attention. The listeners appear to have based their perceptual response to the two speakers disproportionately on the information in the incongruous trials.

This finding obviously clashes with strictly experience-based models of sociophonetic knowledge in which learning is exclusively input-driven. It suggests that exemplar models require a mechanism which allows for the systematic failure of learning to occur, or rather, which allows listeners to actively shape their experience. I suggest that this mechanism is selective attention.

The notion of attention is included among the features of Johnson's (1997) exemplar model of speech perception (following Nosofsky 1988) in the form of *attention weights*. In Johnson's model, a weight parameter allows for variation along some stimulus dimensions to be perceptually emphasized or de-emphasized. As a result of varying attention, some associations between different stimulus dimensions have a greater effect on memory than others. However, besides Johnson's (1997) use of attention weighing to account for some aspects of speaker normalization, the notion has not been made use of in exemplar-based accounts of sociophonetic knowledge even though attention has

been discussed in exemplar models developed to account for non-linguistic behavior in cognitive psychology (e.g., Nosofsky 1986) and social psychology (e.g., Smith and Zárate 1992).

I suggest that attention can account not only for the present results but also for another well-known case of speakers apparently failing to learn associations between social and linguistic variation in their speech community, the Anglo Detroiters studied by Niedzielski (1997, 1999). My account of these data is generally in line with Niedzielski's own interpretation of the effect being caused by linguistic stereotypes (e.g., Niedzielski 2010). However, the discussion here is more close in spirit to Hay, Nolan & Drager's (2006) exemplar-based re-interpretation of the phenomenon and includes a more explicit treatment of the role of selective attention.

In Niedzielski's vowel matching task, which was discussed in Chapter 1, the Detroit listeners who were told that they were listening to a speaker from Michigan did not match the non-standard variant [ʌʊ] of the vowel /aʊ/ in the speech of a fellow Detroiter to a synthetic [ʌʊ]. Instead, they reported hearing the standard variant [aʊ]. The participants in the other condition, who were told

that the speaker they were listening to was from Canada, accurately matched the non-standard variant to the synthetic [ʌʊ]. Niedzielski explained this with reference to Anglo Detroiters' strong belief, or stereotype, that they speak standard English. Canadians, on the other hand were pointed out as speaking with an accent by some of Niedzielski's interviewees, including the use of the raised /aʊ/ variants like [ʌʊ].

As pointed out by Hay, Nolan & Drager (2006), this finding appears to conflict with exemplar-based models. After all, if the Detroit listeners are exposed to the variant [ʌʊ] in the speech of other Michiganders, one would expect them to store exemplars of these variants produced by Michigan speakers. Hay et al. nevertheless offer an account of Niedzielski's results in exemplar terms. Their account rests on the notion of social indexing of exemplars. They argue that the Detroiters do in fact perceive and store the non-standard variants in the speech of fellow Detroiters. However, because of their stereotype which equates Detroit English with Standard English they do not index such tokens as 'Detroit English' or 'Michigan English' but as 'Standard English.' In fact, they have no category of 'Detroit English' or 'Michigan English.' In Niedzielski's

experiment, the listeners in the 'Michigan' condition matched the speaker' [ʌʊ]-like variants to synthetic [aʊ] tokens because the instructions caused them to activate their 'Standard English' variants of /aʊ/. This distribution not only includes Detroit speech but also large amounts of actual standard English [aʊ] variants which the listeners are exposed to, for example, in the media. In Hay et al.'s account this causes the activated exemplar distribution to be dominated by standard, [aʊ]-like variants, and the participants' perception to be skewed in the direction of [aʊ]. The listeners in the 'Canada' condition, on the other hand, responded accurately because their category of 'Canadian English' is accurately dominated by [ʌʊ]-like variants.

One problem with Hay et al.'s account is that it leaves open how social indexing under the influence of stereotypes actually works. What is the mechanism by which listeners fail to associate particular linguistic variants with a social category? I suggest that this mechanism is the notion of attention weighing, as in Johnson's (1997) model. Rather than to say that Detroiters store the variants they hear from other Detroiters as 'Standard English,' it may be more accurate to say that when Detroiters listen to other Detroiters their

stereotype of themselves as standard speakers leads them to pay little or no attention to non-standard features. In this account, as in Hay et al.'s account, Detroiters do perceive and store non-standard features. However, these features may well be indexed as 'Detroit English' or 'Michigan English,' rather than as 'Standard English,' it is only that they are weighed weakly and thus have little influence on the distribution of exemplars activated by the label 'Michigan.' When listening to speakers from Canada, the Detroiters' stereotype of Canadian English as an accented form of English draws their attention to non-standard features. In fact, this may cause them to overestimate the presence and the quality of non-standard features in Canadian English, as reflected in stereotypes of this variety as containing forms like "oot" (*out*) or "aboot" (about), with a hyper-raised onset, as reported by Niedzielski (1997).

One crucial difference between the attention-based account outlined above for Niedzielski's results and the attention-based account of the current results is that the listeners in the current experiment were not guided by a stereotype. Their reason for paying more attention to the Southern variants of /eɪ/ and /ɛ/ in the speech of the younger speaker was simply that it did not

match their experience. Thus, there may be different reasons for the unequal allocation of attention. The question why some situations draw more attention than others is too large to be answered here. The point is merely that differential attention and attention weighing may be the mechanism by which listeners actively steer exemplar-based learning.

In conclusion, experience-based accounts of sociophonetic knowledge such as exemplar models can profit from incorporating the notion of selective attention to phonetic detail. The notion that the storage of novel exemplars is mediated by the allocation of attention to particular types of experience, such as particular sociophonetic variants, while others have a disproportionate impact, accounts for the unexpected learning effect seen in the present experiment. In addition, as I argued above, it sheds light on another known case of dialect perception in which listeners appear to be oblivious to the presence of some variants even though there is abundant evidence in them in a speech community.

References

Abbott, Carl. 1987. The new urban America: Growth and politics in Sunbelt cities. 2nd edition. Chapel Hill: University of North Carolina Press.

Adank, Patti, Roel Smits and Roeland van Hout. 2004. A comparison of vowel normalization procedures for language variation research. *JASA* 116: 3099-3107.

Alwan, Abeer, Philbert Bangayan, Bruce R. Gerratt, Jody Kreiman and Christopher Long. 1995. Analysis by synthesis of pathological voices using the Klatt synthesizer. In: Raymond D. Kent and Martin J. Ball (eds.) *Voice quality measurement.* Singular, pp. 307-325.

Baayen, R. Harald. 2008. Analyzing linguistic data: A practical introduction to statistics. Cambridge: Cambridge University Press.

Bailey, Guy. 1991. Directions of change in Texas English. *Journal of American Culture* 14: 125-134.

Bailey, Guy and Cynthia Bernstein. 1989. Methodology for a Phonological Survey of Texas. *Journal of English Linguistics* 22: 6-16.

Bailey, Guy, Thomas Wikle, and Lori Sand. 1991. The focus of linguistic innovation in Texas. *English World-Wide* 12: 195-214.

Baker, Adam. 2006. Quantifying diphthongs: a statistical technique for distinguishing formant contours. Paper presented at NWAV 35, The Ohio State University.

Bates, D. and D. Sarkar. 2007. lme4: Linear mixed-effects models using s4 classes. Available at: http://r-forge.r-project.org/projects/lme4.

Bigham, Douglas S. 2009. Correlation of the Low-Back Vowel Merger and TRAP-Retraction. *University of Pennsylvania Working Papers in Linguistics* 15.2: 21-31.

Boersma, Paul and David Weenink. 1992-2011. *Praat: doing phonetics by computer*. Available at: http://www.praat.org/

Bowie, David. 2001a. Perception and production in a series of related mergers. In: Ruth M. Brend, Alan K. Melby and Arle R. Lommel (eds.) *LACUS Forum XXVII: Speaking and Comprehending*. Fullerton, California: Linguistic Association of Canada and the United States, pp. 297-305.

Bowie, David. 2001b. Dialect contact and dialect change: The effect of near-mergers. *University of Pennsylvania Working Papers in Linguistics* 7.3: 17-26.

Bybee, Joan. 2001. Phonology and language use. Cambridge: Cambridge University Press.

Clarke, Constance M. and Merrill F. Garrett. 2004. Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America* 116: 3647-3658.

Clopper, Cynthia G., Janet B. Pierrehumbert, and Terrin N. Tamati. 2008. Lexical bias in cross-dialect word recognition in noise. Paper presented at *Laboratory Phonology* 11.

Clopper, Cynthia G. and David B. Pisoni. 2004a. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics* 32: 111-140.

Clopper, Cynthia G. and David B. Pisoni. 2004b. Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change* 16: 31-48.

Clopper, Cynthia G., David B. Pisoni and Adam T. Tierney. 2006. Effects of open-set and closed-set task demands on spoken word recognition. *Journal of the American Academy of Audiology* 17: 331-349.

Drager, Katie. 2005. The influence of social characteristics on speech perception. MA thesis, University of Canterbury.

Drager, Katie. 2009. A sociophonetic ethnography of Selwyn Girls' High. PhD dissertation, University of Canterbury.

Drager, Katie. 2011. Speaker age and vowel variation. *Language and Speech* 54: 99-121.

Duffy, Susan A., and David B. Pisoni. 1992. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech* 35: 351–389.

Feagin, Crawford. 1987. A closer look at the Southern drawl: variation taken to the extremes. In: Keith M. Denning, Sharon Inkelas, Faye C. McNair-Knox and John. R. Rickford (eds.) *Variation in Language. NWAV-XV at Stanford: Proceedings of the Fifteenth Annual Conference on New Ways of Analyzing Variation.* Stanford: Department of Linguistics, pp. 137-50.

Flanigan, Beverly Olson and Franklin Paul Norris. 2000. Cross-dialect comprehension as evidence for boundary mapping: Perceptions of the speech of Southeastern Ohio. *Language Variation and Change* 12: 175-201.

Foreman, Christina. 2000. Identification of African-American English from prosodic cues. *Texas Linguistic Forum* 43: 57-66.

Foulkes, Paul, Gerard J. Docherty, Ghada Khattab, and Malcah Yaeger-Dror. 2010. Sound judgments: Perception of indexical features in children's speech. In: Dennis R. Preston and Nancy A. Niedzielski (eds.) *A reader in sociophonetics*. Berlin: Mouton de Gruyter.

Glidden, Catherine M. and Peter Assmann. 2004. Effects of visual gender and frequency shifts on vowel category judgments. *Acoustic Research Letters Online* 5: 132-138.

Gentry, Elizabeth. 2006. Hovering between the South and the West: Houston's merged dialect. Paper presented at NWAV 35.

Goldinger, Stephen D. 1997. Words and voices. Perception and production in an episodic lexicon. In: K. Johnson and J. Mullennix (eds.) *Talker variability in speech processing*. London: Academic Press, pp. 33-66.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105: 251-279.

Goldstone, Robert. L. 1998. Perceptual Learning. *Annual Review of Psychology* 49: 585-612.

Graff, David, William Labov, and Wendell A. Harris. 1986. Testing listeners' reactions to phonological markers of ethnic identity: A new method for sociolinguistic research. In: D. Sankoff (ed.) *Diversity and Diachrony*. Amsterdam/ Philadelphia: Benjamins, pp. 45-58.

Hagiwara, Robert. 1997. Dialect variation and formant frequency: The American English vowels revisited. *JASA* 102: 655–658.

Hay, Jennifer and Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48: 865-892.

Hay, Jennifer, Aaron Nolan, and Katie Drager. 2006. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23: 351-379.

Hay, Jennifer, Paul Warren, and Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34: 458-484.

Hillenbrand, James M., Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of Americaec* 97: 3099-3111.

Janson, Tore. 1983. Sound change in perception and production. *Language* 59: 18-34.

Janson, Tore. 1986. Sound change in perception: an experiment. In: J. Ohala and J. Jaeger (eds.) *Experimental phonology*. Orlando: Academic Press, pp. 253-260.

Janson, Tore and Richard Schulman. 1983. Non-distinctive features and their use. *Journal of Linguistics* 19: 321-336.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359-383.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In: Keith Johnson and John W. Mullennix (eds.) *Talker variability in speech processing*. San Diego: Academic Press, pp. 145-165.

Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34: 485-499.

Johnson, Keith and John W. Mullennix. 1997. Talker variability in speech processing. San Diego: Academic Press.

Johnson, Keith, Elizabeth A. Strand, and Mariapaola D'Imperio. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27: 359-384.

Koops, Christian, Elizabeth Gentry, and Andrew Pantos. 2008. The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics* 34.2: 93-101.

Kraljic, Tanya, and Arthur G. Samuel. 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56: 1-15.

Labov, William. 1966. The social stratification of English in New York City. Washington, DC: Center for Applied Linguistics.

Labov, William. 1994. Principles of Linguistic Change. Vol. I: Internal Factors. Oxford: Basil Blackwell.

Labov, William. 2001. Principles of Linguistic Change. Vol. II: Social Factors. Oxford: Basil Blackwell.

Labov, William. 2006. A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics* 34: 500-515.

Labov, William, and Sharon Ash. 1997. Understanding Birmingham. In: C. Bernstein, T. Nunnally, and R. Sabino (eds.) *Language variety in the South revisited.* Tuscaloosa: University of Alabama Press, pp. 508-573.

Labov, William, Sharon Ash, and Charles Boberg. 2006. The Atlas of North American English. Phonetics, Phonology and Sound Change. New York: Mouton de Gruyter.

Labov, William, Mark Karen, and Corey Miller. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3: 33-74.

Labov, William, Malcah Yaeger, and Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress*. 2 Vols. Philadelphia: U.S. Regional Survey.

Lambert, Wallace E., R. C. Hogson, R. C. Gardner, and S. Fillenbaum. 1960. Evaluative reactions to spoken languages. *Journal of Abnormal and Social Psychology* 60: 44-51.

Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *JASA* 49: 606-608.

Luce, Paul A. and Emily A. Lyons. 1998. Specificity of memory representations for spoken words. *Memory and Cognition* 26: 708-715.

Luce, Paul A., Conor T. McLennan and Jan Charles-Luce. 2003. Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In: Jeffrey S. Bowers and Chad J. Marsolek (eds.) *Rethinking Implicit Memory*. Oxford University Press, pp. 197-214.

Marslen-Wilson, William & Lorraine Komisarjevsky Tyler. 1980. The temporal structure of spoken language understanding. *Cognition* 8: 1-71.

Maye, Jessica, Richard N. Aslin, and Michael K. Tanenhaus. 2008. The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science* 32: 543-562.

McDougall, Kirsty and Francis Nolan. 2007. Discrimination of speakers using the formant dynamics of /u:/ in British English. *ICPhS XVI, Saarbrücken, 6-10 August 2007*, pp. 1825-1828.

McLennan, Conor T. and Paul A. Luce. 2005. Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31:306-321

Nearey, Terrance Michael. 1977. Phonetic feature systems for vowels. Ph.D. Dissertation, University of Alberta. Reprinted 1978 by the Indiana University Linguistics Club.

Ní Chasaide, Ailbhe and Christer Gobl. 1993. Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech* 36: 303-330.

Niedzielski, Nancy. 1997. The effect of social information on the phonetic perception of sociolinguistic variables. PhD dissertation, University of California, Santa Barbara.

Niedzielski, Nancy. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18: 62-85.

Niedzielski, Nancy. 2006. Language, perception, and 'remarkable times': using demography in sociolinguistic research. Poster presentation at NWAV 35, Columbus, Ohio, November 9-12, 2006.

Niedzielski, Nancy A. 2010. Linguistic security, ideology, and vowel perception. In: Dennis Preston and Nancy Niedzielski (eds.) *A Reader in Sociophonetics*. New York: De Gruyter Mouton, pp. 253-264.

Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 15: 39-57.

Nosofsky, Robert M. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 700-708.

Nunnally, and R. Sabino (eds.) *Language variety in the South revisited.* Tuscaloosa: University of Alabama Press, pp. 508-573.

Nycz, Jennifer, and Paul De Decker. 2006. A new way of analyzing vowels: Comparing formant contours using Smoothing Spline ANOVA. Poster presented at NWAV 35, The Ohio State University.

Pantos, Andrew. 2006. Redefining the South. Teenage Houstonians and the Southern Shift. Talk at NWAV 35.

Peterson, Gordon E. and H. L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 175-184.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In: Joan Bybee and Paul Hopper (eds.) *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins, pp. 137-157.

Pisoni, David B. 1993. Long-term memory in speech perception: Some new findings on talker variability, speaking rate, and perceptual learning. *Speech Communication* 13: 109-25.

Pisoni, David B. 1997. Perception of synthetic speech. In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (eds.) *Progress in Speech Synthesis.* New York: Springer, pp. 541-560.

Pisoni, David B., H. C. Nusbaum, P. A. Luce and L. M. Slowiaczek. 1985. Speech perception, word recognition, and the structure of the lexicon. *Speech Communication* 4, 75–95.

Plichta, Bartłomiej and Dennis R. Preston. 2005. The /ay/s have it: the perception of /ay/ as a North-South stereotype in US English. *Acta Linguistica Hafniensia* 37: 107-130.

Preston, Dennis R. 2005. Belle's body just caught the fit gnat: The perception of Northern Cities shifted vowels by local speakers. *University of Pennsylvania Working Papers in Linguistics* 11.2: 133-146.

Purnell, Thomas C., William Idsardi and John Baugh. 1999. Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology* 18: 10-30.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org.

Smith, Eliot R. and Michael A. Zárate. 1992. Exemplar-based model of social judgment. *Psychological Review* 99: 3-21.

Staum, Laura. 2008. Experimental investigations of sociolinguistic knowledge. PhD dissertation, Stanford University.

Strand, Elizabeth A. 2000. Gender stereotype effects in speech processing. PhD dissertation. The Ohio State University.

Strand, Elizabeth A., and Keith Johnson. 1996. Gradient and visual speaker normalization in the perception of fricatives. In: D. Gibbon (ed.) *Natural language processing and speech technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996*. Berlin: Mouton, pp. 14-26.

Sumner, Meghan and Arthur G. Samuel. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language* 60: 487-501.

Thomas, Erik R. 1997. A rural/metropolitan split in the speech of Texas Anglos. *Language Variation and Change* 9: 309-332.

Thomas, Erik R. 2003. Secrets revealed by Southern vowel shifting. *American Speech* 78: 150-170.

Thomas, Erik R. 2007. Phonological and phonetic characteristics of African American Vernacular English. *Language and Linguistics Compass* 1: 450-475.

Thomas, Erik R., Norman J. Lass and Jeannine Carpenter. 2010. Identification of African American Speech. In: Dennis R. Preston and Nancy Niedzielski (eds.) *A Reader in Sociophonetics*. Cambridge, UK: Cambridge University Press.

Thomas, Erik R. and Jeffrey Reaser. 2004. Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of Sociolinguistics* 8: 54-87.

Tillery, Jan and Guy Bailey. 2003. Urbanization and the evolution of Southern American English. In: S. Nagle and S. Sanders (eds.) *English in the Southern United States.* Cambridge: Cambridge University Press, pp. 159-171.

van Bezooijen, Renée and Charlotte Gooskens. 1999. Identification of language varieties. The contribution of different linguistic levels. *Journal of Language and Social Psychology* 18: 31-48.

Walker, Abby. 2007. The effect of phonetic detail on perceived speaker age and social class. Proceedings of ICPhS XVI, Saarbrücken, 6-10 August 2007, pp. 1453-1456.

Warren, Paul, Jennifer Hay, and Brynmor Thomas. 2007. The loci of sound change effects in recognition and perception. In: Jennifer Cole and José I. Hualde (eds.) *Laboratory Phonology 9*. New York: Mouton de Gruyter, pp. 87-111.

Weenink, David. The KlattGrid speech synthesizer. Paper presented at Interspeech 2009, Brighton, UK, 6-10 September 2009.

Wetzell, Brett. 2000. Rhythm, dialects, and the Southern drawl. MA thesis, North Carolina State University.

Williams, Angie, Peter Garrett, and Nikolas Coupland. 1999. Dialect recognition. In: Dennis R. Preston (ed.) *Handbook of perceptual dialectology*. Vol. 1. Philadelphia: Benjamins, pp. 345-358.

Willis, Clodius. 1972. Perception of vowel phonemes in Fort Erie, Ontario, Canada, and Buffalo, New York: an application of synthetic vowel categorization tests to dialectology. *Journal of Speech and Hearing Research* 15: 246-255.

Yaeger-Dror, Malcah and Erik R. Thomas (eds.) 2010. *African American English Speakers and their Participation in Local Sound Changes: A Comparative Study*. Duke University Press.